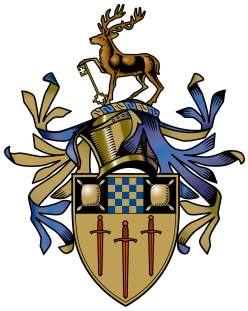
Intelligent Video Understanding: Self-Supervised and Multimodal Learning

Mona Ahmadian

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey



Faculty of Arts, Business and Social Sciences
University of Surrey

Principal Supervisor: Dr. Andrew Gilbert
Co-supervisor: Dr. Frank Guerin

October 2025

Abstract

In the rapidly evolving world of visual media, accurately recognising and localising complex actions in untrimmed, real-world videos remains a fundamental challenge. It demands modelling fine-grained motion dynamics, capturing long-range temporal dependencies, and interpreting high-level semantics across multiple modalities. Despite remarkable progress in visual representation learning, existing methods often struggle to reason over complex motion, align visual features with linguistic meaning, and integrate complementary audio cues for context-aware understanding. This thesis addresses these challenges through a unified investigation of motion reasoning, language-grounded semantics, and multimodal integration, aiming to enhance the interpretability and precision of video understanding systems.

First, we propose MOFO (MOtion FOcused Self-Supervision for Video Understanding), a framework that explicitly models motion dynamics during self-supervised pretraining and finetuning for action recognition. MOFO automatically detects motion-sensitive regions using a motion map derived from optical flow derivatives, highlighting motion boundaries while reducing the impact of camera movement and background noise. A motion-guided masking strategy focuses learning on dynamic, action-relevant areas, encouraging the model to capture motion cues rather than static appearance. During finetuning, a multi-cross attention mechanism fuses embeddings from inside and outside the detected motion regions, improving temporal reasoning and contextual understanding. By integrating motion guidance across both pretraining and finetuning, MOFO produces interpretable, motion-aware representations and significantly enhances self-supervised action recognition performance.

Next, we introduce FILS (Self-Supervised Video Feature Prediction in Semantic Language Space), a framework that enhances video representation learning by predicting features within a language-aligned semantic space. FILS first constructs this semantic space through contrastive learning between motion-relevant video patches and automatically generated textual descriptions, aligning dynamic visual regions with their corresponding language embeddings. It then performs feature prediction for the masked video patches within the learnt language space, allowing the model to capture high-level semantic meaning without pixel reconstruction. By combining motion-aware contrastive learning with language-guided feature prediction, FILS learns interpretable and transferable video representations that bridge visual motion dynamics with linguistic understanding.

Finally, we present DEL (Dense Event Localisation for Multimodal Audio-Visual Understanding), a multimodal framework for precise and fine-grained event detection in long, untrimmed videos where events of different durations may overlap and occur asynchronously. Unlike MOFO and FILS, which focus on self-supervised representation learning, DEL employs supervised multimodal learning to model dense temporal structures and cross-modal dependencies in realistic video environments. It integrates visual and audio modalities through an adaptive cross-modal attention mechanism that aligns asynchronous cues and preserves temporal coherence. To improve feature discrimination and robustness, DEL introduces a score-based dual contrastive learning strategy that enhances intra- and inter-modal consistency. A hierarchical temporal fusion module further aggregates information across multiple temporal scales, enabling the model to capture both short-term motion details and long-range contextual dependencies. In combination, these components allow DEL to accurately detect overlapping and asynchronous audio–visual events while remaining computationally efficient.

Together, these studies form a coherent progression toward comprehensive video understanding, evolving from motion perception to semantic reasoning and finally to multimodal temporal analysis. The results show that focusing on motion leads to stronger and more interpretable representations, grounding visual features in language enhances semantic understanding, and integrating audio and visual modalities enables precise and temporally coherent event localisation. Overall, these contributions advance unified and semantically grounded approaches to video understanding, bridging the gap between visual perception and high-level reasoning.

Key words: Deep Learning, Machine Learning, Computer Vision, Video Understanding, Self-Supervised Learning, Motion-Aware Action Recognition, Audio-Visual Event Localisation

Email: m.ahmadian@surrey.ac.uk
WWW: https://github.com/Moohnai

Acknowledgements

I am deeply grateful to everyone who has supported and encouraged me throughout this PhD journey.

I would like to express my sincere and deepest gratitude to my supervisors, Dr.Andrew Gilbert and Dr.Frank Guerin, for their invaluable guidance, encouragement, and support throughout my PhD journey. Their expertise, patience, and insightful feedback have greatly shaped this research and my development as a scholar. Your commitment to your students is truly inspiring, and I feel deeply fortunate to have had the opportunity to learn and work under your supervision.

I am deeply thankful to my mother and sister, whose love and encouragement have supported me even across the distance. Their constant belief in me has given me strength and motivation throughout this journey.

Most of all, I owe my deepest thanks to my husband, Amir, my greatest supporter and best friend. Your endless love, patience, and belief in me have carried me through every challenge and made this journey possible. This accomplishment belongs to you as much as it does to me.

Declaration

This thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images or text resulting from the work of others (whether published or unpublished, and including any content generated by a deep learning/artificial intelligence tool) are fully identified as such within the work and attributed to their originator in the text, bibliography or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification. I agree that the University has the right to submit my work to the plagiarism detection service TurnitinUK for originality checks. Whether or not drafts have been so-assessed, the University reserves the right to require an electronic version of the final document (as submitted) for assessment as above.

The studies and findings presented in this thesis are also presented in the following manuscripts:

1. M. Ahmadian, F. Guerin, and A. Gilbert. "MOFO: MOtion FOcused Self-Supervision for Video Understanding," *In Proceedings Self-Supervised Learning Workshop- Theory and Practice (NeurIPS), 2023*. **(Chapter 3)**

2. M. Ahmadian, F. Guerin, and A. Gilbert. "FILS: Self-Supervised Video Feature Prediction In Semantic Language Space," *In Proceedings of 35th British Machine Vision Conference (BMVC), 2024*. **(Chapter 4)**

3. M. Ahmadian, A. Shirian, F. Guerin, and A. Gilbert. "DEL: Dense Event Localisation for Multi-modal Audio-Visual Understanding," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) sWorkshop on What is Next in Multimodal Foundation Models!(MMFM4) / Under Review at WACV 2026*. **(Chapter 5)**

Signed: Mona Ahmadian

Date: 10/2025

Contents

Nomenclature

Acronyms

SSL	Self-Supervised Learning
MAE	Masked Autoencoder
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
ViT	Vision Transformer
MCA	Multi-head Cross Attention
IoU	Intersection over Union
MSE	Mean Squared Error
TV-L1	Total Variation L1 (optical flow algorithm)
GradCAM	Gradient-weighted Class Activation Mapping
FC	Fully Connected (layer)
CLIP	Contrastive Language-Image Pretraining
ActCLIP	Patch-wise video–text contrastive learning within detected action areas
FP	Feature Prediction
EMA	Exponential Moving Average
LLM	Large Language Model
mAP	Mean Average Precision
AdamW	Adaptive Moment Estimation with Weight Decay

FlashAttention	Fast Memory-Efficient Attention Mechanism
TAL	Temporal Action Localisation
AVEL	Audio-Visual Event Localisation
FPN	Feature Pyramid Network
AAC	Adaptive Attention for Cross-Modal Alignment
SCL	Score-based Contrastive Learning
PAN	Path Aggregation Network
APM	Adaptive Pooling Module
CLS	Classification Token
RGB	Red, Green, Blue (Colour Space)
tIoU	Temporal Intersection over Union
MHA	Multi-Head Attention
I3D	Inflated 3D Convolutional Network
VGGish	VGG-based Audio Feature Extractor pretrained on AudioSet
TP / FP	True Positive / False Positive
GT	Ground Truth

Symbols

Introduced in Chapter 3

T	Total number of frames in a video sequence.
H, W	Height and width of a video frame in pixels.
f_i	Optical flow, pixel-level motion between frames i and $i+1$.
u_i, v_i	Horizontal and vertical optical flow components between frames i and $i+1$.
m_i	Motion map of frame i highlighting local motion boundaries.
$G_{x,y}$	Gaussian smoothing operator applied over the motion map to remove high-frequency noise
σ	Standard deviation of the Gaussian kernel controlling smoothing.
$\hat{m}_i$	Smoothed motion map obtained after Gaussian filtering.
$\mathbf{p}_i$	3D video tube or patch extracted from the input video; $\mathbf{p}_i \in \mathbb{R}^{H_t \times W_t \times T_t}$.
H_t, W_t, T_t	Height, width, and temporal depth of each tube patch.
N	Total number of non-overlapping tube patches.
$N_{\text{inner}}, N_{\text{outer}}$	Embeddings of inner and outer motion areas.
$\mathbf{e}^{\text{inner}}, \mathbf{e}^{\text{outer}}$	Fused embedding combining information from both motion (inner) and context (outer) regions through Multi-Cross Attention
Q, K, V	Query, key, and value matrices.
$d_k, d_v, d_{\text{model}}$	Dimensions of key, value, and model embeddings.
W_i^Q, W_i^K, W_i^V	Learnable projection matrices for the i^{th} attention head.
W^O	Output projection matrix after concatenating attention heads.
head_i	Output of the i^{th} attention head in multi-head cross-attention.

h	Number of attention heads in multi-cross attention.
$\mathbf{e}^{\text{fused}}$	Fused embedding after multi-cross attention.
$\mathbf{y}_n, \hat{\mathbf{y}}_n$	Ground-truth and predicted label vectors for the n^{th} video clip.

Introduced in Chapter 4

V	Input video sequence composed of multiple frames.
V_u	Unmasked patches of the input video used by the student encoder.
V_m	Masked patches of the input video used by the teacher encoder.
f^u	Feature embedding of unmasked patches from the student encoder.
f^m	Feature embedding of masked patches from the teacher encoder.
p^m	Predicted feature embeddings of masked patches generated by the predictor.
N_u	Number of unmasked (visible) patches in the video.
N_m	Number of masked patches in the video.
N	Total number of patches in the video, where $N = N_u + N_m$.
m, u	Indices referring to masked and unmasked patches, respectively.
a	Index representing patches within the detected action area.
N_a	Number of patches within the detected action area.
$\bar{f}$	Mean-pooled feature of patches within the detected action area.
z^V	Normalised projected video feature in the shared semantic space.
z^T	Normalised projected text feature in the shared semantic space.
$\theta(\cdot)$	Projection head (mapping function) that transforms features from the video space to the language space.
$\|\cdot\|$	Normalisation operation.
h	Encoded text representation produced by the text encoder.
T	Input textual sequence or caption associated with the video.
Enc$V$Student$(\cdot)$	Student video encoder that processes unmasked video patches.
Enc$V$Teacher$(\cdot)$	Teacher video encoder that processes masked video patches.

Enc$L(\cdot)$	Text encoder mapping input text to latent representations.
Predictor$(\cdot)$	Transformer-based decoder, predicting masked patch features.
Δ	Teacher model parameters updated via exponential moving average.
θ	Student model parameters.
τ	Momentum coefficient in the exponential moving average (EMA) update rule.
B	Batch size during training.
i, j	Indices representing samples within a batch.
σ	Learnable temperature parameter in contrastive loss.
$\langle \cdot, \cdot \rangle$	Cosine similarity operation between two feature vectors.
L_{V2T}	Video-to-text contrastive loss.
L_{T2V}	Text-to-video contrastive loss.
λ_1, λ_2	Weighting coefficients for balancing $L_{ActCLIP}$ and L_{FP} in total loss.
$D(\cdot, \cdot)$	Distance metric (e.g., L1 distance) used for computing feature prediction loss.
N_m	Number of masked patches in the video input.
$[MASK]$	Learnable token used by the predictor during reconstruction.
$\tilde{p}_i^m$	Normalised mapped predicted patch representation in language space.
$\tilde{g}_i^m$	Normalised mapped target (ground-truth) patch representation in language space.

Introduced in Chapter 5

$\mathbb{S}$	Set of paired audio-visual segments $\{(\mathbf{V_t}, \mathbf{A_t})\}_{t=1}^{T}$.
$\mathbf{V_t}$	Visual feature representation at time t.
$\mathbf{A_t}$	Audio feature representation at time t.

T	Total number of temporal segments in a video.
$\mathcal{G}$	Ground-truth annotation set $\{g_n = (\tau_{start,n}, \tau_{end,n}, \lambda_n)\}_{n=1}^{N}$.
$\tau_{start,n}, \tau_{end,n}$	Start and end timestamps of the n^{th} event.
λ_n	Event class label of the n^{th} event.
Λ	Set of all event categories.
N	Total number of annotated events in the video.
$\hat{\mathbb{S}}$	Predicted event set $\{\hat{s}_t = (\delta_{start,t}, \delta_{end,t}, q(y_t))\}_{t=1}^{T}$.
$q(y_t)$	Predicted class probability distribution over $\lvert\Lambda\rvert$ classes at time t.
$\delta_{start,t}, \delta_{end,t}$	Predicted temporal offsets from time t to event boundaries.
$\hat{\lambda}_t$	Predicted event label at time t.
$\mathbf{M}$	Event-aligned attention mask guiding cross-modal interactions.
L_v, L_a	Lengths of video and audio feature sequences, respectively.
$\mathbf{X}$	Concatenated feature sequence
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Query, key, and value matrices.
$\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$	Learnable projection matrices $\in \mathbb{R}^{d \times d}$.
d	Feature embedding dimension used for query, key, and value projections in the attention mechanism.
$\text{aat}_{i,j}$	Adaptive attention coefficient between features i and j.
$m_{i,j}$	Binary entry of the event-aligned attention mask M between features i and j.
s_t	Binary score indicating event presence at time t.
c_t	Predicted event category for feature t.
$I_{PV}, I_{HNV}, I_{PA}, I_{HNA}$	Positive and hard-negative feature sets for video and audio modalities.
$\ell(z, z^+, z^-)$	Contrastive loss function.
τ	Temperature parameter in contrastive loss.
$\mathcal{L}_{inter}$	Inter-sample contrastive loss.
$\mathcal{L}_{intra}$	Intra-sample contrastive loss.
$\mathcal{L}_{score}$	Binary cross-entropy loss for score prediction.
$\mathcal{L}_{cls}$	Cross-entropy classification loss.
$\mathcal{L}_{reg}$	Smooth L1 loss for temporal boundary regression.

$\mathcal{L}_{DEL}$	Overall training objective for DEL.
$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$	Weighting coefficients for loss terms.
V_l, A_l	Multiscale visual and audio features at pyramid level l.
C_V, C_A	Number of embedding channels in visual and audio modalities.
V_l', A_l'	Updated features after modality-guided adaptation.
σ	Sigmoid activation function.
$\max(\cdot)$	Max pooling operator applied across feature channels.
$\tilde{V}, \tilde{A}$	Pooled multiscale visual and audio tokens.
$\mathrm{MHA}(\cdot)$	Multi-Head Attention operation.
L	Number of pyramid levels in the path aggregation network.
P	Precision metric for event detection.
AP	Average precision across a class.
mAP	Mean Average Precision across all classes.

List of Figures

List of Tables

Chapter 1

Introduction

The rapid growth of video data across diverse platforms and devices has changed how we share information, interact with computers, and communicate through multimedia. From wearable cameras and mobile phones to autonomous vehicles and surveillance systems, modern society generates vast amounts of dynamic, multimodal visual content every second [118, 88, 181].

Understanding video content by identifying its actions, events, and interactions lies at the core of intelligent video analysis. Unlike static image analysis, which relies primarily on spatial semantics, video understanding requires reasoning over time, motion, and multiple modalities [223, 239, 4].

Human perception provides a natural example of what intelligent understanding looks like. We constantly combine information from sight, sound, touch, motion, and other senses to form a coherent picture of the world, recognising patterns, anticipating changes, and interpreting meaning. This multimodal and temporal process depends on individual cues and how they evolve over time. Our ability to understand movement, context, and continuity enables us to interpret everyday situations such as conversations or physical activity. Crucially, this perception remains stable under changing conditions; we can still recognise people or situations under different lighting, viewpoints, or sounds.

Building such perceptual abilities in machines is the core goal of intelligent video understanding. Videos capture visual appearance, change, causality, and interaction over time. Sound and language often enrich them with contextual depth and analytical complexity. Achieving effective

video understanding requires learning how appearance, motion, and audio cues interact and relate to real-world meaning, modelling both spatial and temporal structures to capture what happens within each frame and how those moments evolve [68, 270, 278]. This highlights the importance of integrating diverse cues and maintaining flexible, generalisable representations [84, 200, 229, 265].

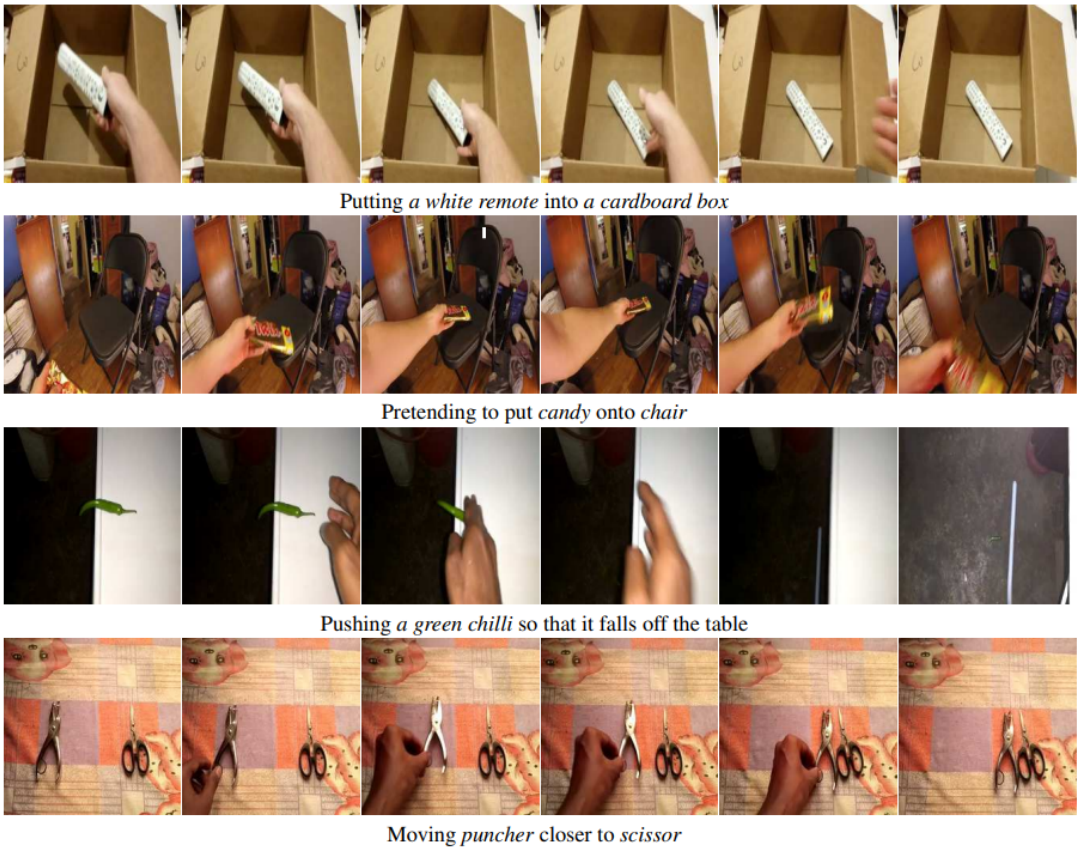


Figure 1.1: Examples from the Something-Something dataset [87] illustrating actions that are visually similar in individual frames but distinguishable only through temporal evolution.

Despite significant advances in deep learning, video understanding remains challenging due to the complexity and inconsistency of video data [87, 47]. Useful information may appear briefly, while much of the content is redundant or irrelevant. This challenge is illustrated in fig. 1.1, adapted from the Something-Something dataset [87], where subtle temporal variations across visually similar frames are crucial for distinguishing semantically different actions. Since annotating large video datasets is costly and often impractical, it is crucial to develop methods that learn directly from data, discovering meaningful structure without heavy human supervision [98, 32, 238]. Self-supervised learning enables models to capture motion and temporal

relationships directly from the data by exploiting the inherent structure of video, recognising how visual information evolves over time without manual labels [249, 284, 71]. However, these methods often lack deeper semantic understanding. To achieve more human-like interpretation, models must link what they see and hear to higher-level meaning, associating actions and sounds with intentions, objects, and context [4, 200, 322]. Specific forms of understanding still require supervision. Complex multimodal videos benefit from annotated data to guide cross-modal alignment and refine temporal localisation [237, 273]. Supervised learning helps capture both global structure and fine-grained, context-specific details essential for accurate event understanding. Beyond these learning challenges, efficiency and accessibility remain major concerns, as training large-scale video models demands substantial computational resources [321, 48, 252, 66]. Video understanding has wide-ranging potential across healthcare, robotics, education, and public safety, where intelligent analysis can support diagnosis, enhance monitoring, and improve decision-making. To realise this potential, future systems must be accurate and semantically aware, efficient, adaptable, and explainable.

This thesis addresses the challenges of efficient, interpretable, and multimodal video understanding by developing advanced deep learning approaches. The research progresses from self-supervised learning for motion-aware representation learning to language-guided semantic modelling and multimodal integration for complex event analysis. Together, these studies establish a unified direction toward learning from unlabelled data, enriching visual understanding with semantic context, and achieving fine-grained, multimodal reasoning in real-world videos. To summarise these challenges, fig. 1.2 provides a conceptual overview of the key components involved in video understanding.

1.1 Applications

The following are some current applications of deep learning-based video understanding in areas such as healthcare, robotics, multimedia, and human–computer interaction, where systems interpret human behaviour, environmental context, and multimodal information.

Egocentric and First-Person Video Understanding

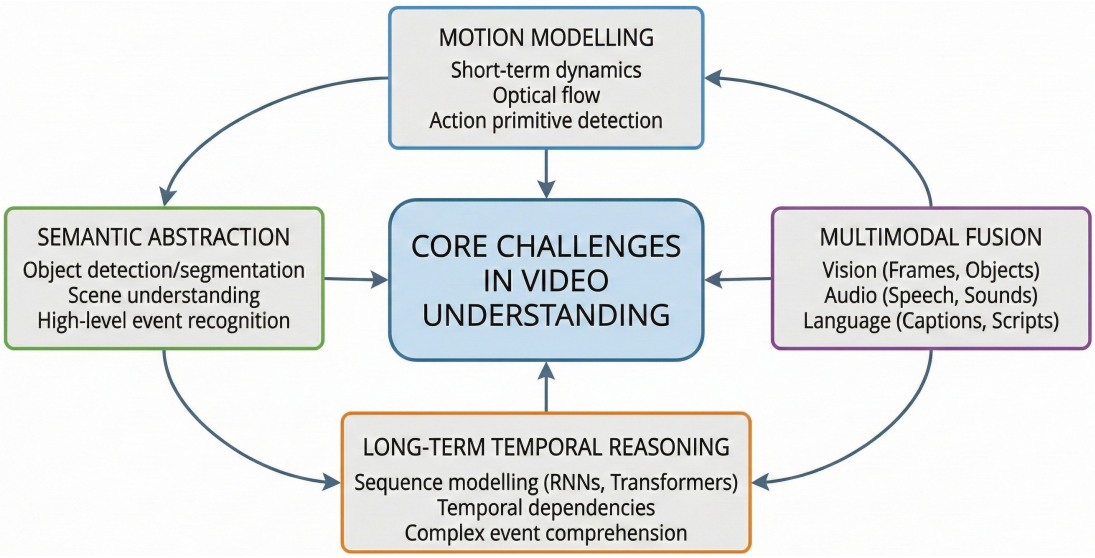


Figure 1.2: Conceptual overview of the core challenges in video understanding.

Egocentric or first-person video analysis has gained importance with the spread of wearable cameras, mobile devices, augmented reality, and robot-mounted sensors [44, 88, 139]. Unlike traditional third-person views, these videos are often recorded with head- or chest-mounted cameras, capturing the world from the actor's own perspective. This first-person view reveals cues such as gaze direction, hand–object interactions, and task intent, providing insight into how actions are experienced and closely reflecting human perception rather than merely observation.

Egocentric video understanding has broad practical applications in assistive technologies, training, robotics, and augmented reality. Wearable cameras can monitor daily activities, support rehabilitation, and assist in surgical or industrial skill assessment. In robotics, egocentric perception links human vision with robotic control, while in augmented reality, it enables context-aware interfaces that adapt to users' actions. However, challenges such as limited field of view, frequent occlusion by hands or objects, and unstable camera motion require models that robustly capture motion dynamics and temporal dependencies. Self-supervised learning is particularly well-suited to egocentric video, where large-scale annotation is costly and impractical. By leveraging motion patterns and contextual cues without explicit supervision, such systems can identify key actions, recognise object manipulations, and segment daily routines [62, 95], ultimately enhancing quality of life for older adults, supporting rehabilitation, and providing cognitive assistance for individuals with memory impairments.

Robotics and Embodied Intelligence

In robotics, egocentric video understanding forms a crucial foundation for embodied perception, learning from demonstration, and human–robot collaboration [213, 201]. Viewing the world from the agent's visual perspective allows robots to better interpret their surroundings, learn from human actions, and generalise across activities. Understanding human intentions and anticipating future actions, such as reaching for a tool or moving an object, improves safety and coordination in shared workspaces. Integrating multimodal cues, including visual, auditory, and motion information, further enhances situational awareness and collaborative performance [111, 196].

Industrial Process Monitoring and Skill Evaluation

Intelligent video systems are increasingly used in industrial and educational settings to analyse task execution and procedural performance [213, 201]. By capturing fine-grained temporal structure and motion dynamics, such systems can automatically assess the quality and completeness of complex workflows. For example, monitoring assembly lines or surgical procedures requires identifying task progression, detecting deviations, and ensuring adherence to safety or operational protocols.

Multimedia Analytics and Content Understanding

In the era of massive online video data, efficient organisation and retrieval of information have become major challenges. Video understanding models that capture semantic meaning and temporal context enable large-scale content categorisation, automatic captioning, and intelligent recommendation [229, 181, 200, 132, 75, 96]. Such capabilities are valuable in media production, digital archiving, and education, where automatic summarisation or search by description can streamline access to vast video repositories. Associating visual events with textual concepts also enhances accessibility for users with visual or cognitive impairments.

Safety, Surveillance, and Public Environments

In safety-critical and public environments, intelligent systems must continuously interpret activities to detect significant events and maintain situational awareness. Advances in video understanding enable more adaptive and context-aware applications such as automated surveillance, crowd analysis, traffic monitoring, and emergency response, where precise temporal and contextual modelling can significantly improve decision-making and safety outcomes [272, 243, 121].

Modelling temporal dependencies helps these systems identify relevant behaviours even under occlusion or partial visibility, improving reliability and safety outcomes.

Healthcare and Well-Being

Video understanding plays an increasingly important role in healthcare and well-being, where analysing patterns of human movement and interaction supports a range of clinical and assistive applications [49, 243, 146]. It enables the detection of abnormal behaviours, assessment of rehabilitation progress, and continuous monitoring for preventive care and early diagnosis. Advances in self-supervised and efficient learning have made these systems more practical by reducing annotation costs and enabling deployment on lightweight devices across hospitals, rehabilitation centres, and home-care environments. Ultimately, video understanding contributes to proactive, accessible, and human-centred healthcare by enhancing automated analysis of human activity for improved well-being and quality of life.

1.2 Challenges

While video understanding has shown remarkable progress across diverse real-world applications, several fundamental challenges remain in learning semantic representations without manual annotation, capturing long-term temporal dependencies, and achieving efficient and coherent multimodal fusion.

1.2.1 Focusing on Meaningful and Dynamic Regions

One of the continuing challenges in video understanding is enabling models to effectively focus on the most informative and dynamic regions within a scene [216, 54]. Unlike static images, videos contain high temporal redundancy, as consecutive frames often repeat similar visual content. The most informative cues for understanding a sequence typically arise from motion, interaction, or subtle temporal change, while much of the data consists of irrelevant background motion or static content [87, 47]. Identifying and prioritising these discriminative spatiotemporal regions is therefore crucial for both efficiency and accuracy.

Motion is a defining characteristic of video data, capturing change and continuity over time [223, 239]. However, effectively leveraging motion without supervision remains difficult. Many self-

supervised methods rely on frame-level reconstruction or global contrastive objectives that overlook local dynamics [98, 32]. Distinguishing meaningful motion, such as object interaction or human gestures, from irrelevant movement caused by camera shifts or background activity remains inherently challenging in the absence of annotated labels [250, 54].

To address this limitation, recent research has explored motion-aware mechanisms, including optical flow estimation, temporal attention, and region-level contrastive learning, to guide representation learning toward dynamic and semantically relevant areas [107, 63, 279]. Focusing on motion-salient regions enables models to form more compact and discriminative embeddings that capture the essence of human actions and object interactions. This challenge becomes even more critical in self-supervised settings, where models must infer temporal saliency directly from data and learn how relevant motion emerges and evolves over time, without any explicit labels to indicate which regions contain meaningful information [292, 283]. Developing frameworks that automatically detect and focus on meaningful motion regions remains essential for building efficient and interpretable video understanding systems.

1.2.2 Learning Semantic Representations Without Labels

One of the most persistent challenges in video understanding is learning meaningful and transferable semantic representations without extensive manual annotation [98, 100, 229, 18]. Creating temporally dense video labels is costly and inconsistent, as it requires defining action boundaries and handling overlapping activities. Predefined label sets also oversimplify the diversity of real-world motion, limiting generalisation to unseen environments or actions. To address these issues, recent research has focused on self-supervised learning, which leverages the inherent structure of video data to learn from pretext tasks such as predicting temporal order, aligning modalities, or reconstructing masked regions [284, 71, 4, 190, 238].

Despite this progress, most self-supervised methods still struggle to move beyond low-level visual and motion cues toward deeper semantic understanding [100, 5]. Techniques like masked reconstruction and contrastive learning often capture how things look or move rather than what they mean. For instance, a model might learn to predict hand movement or background dynamics but fail to infer that the sequence depicts a specific event, such as preparing food or interacting with another person. Achieving a higher level of abstraction requires bridging visual

perception with conceptual understanding, allowing models to associate visual observations with linguistic meaning [229, 171, 182]. Recent work has explored aligning video features with language embeddings to create richer and more interpretable representations that generalise across domains [100, 205]. Nevertheless, learning semantic representations without labels remains an open challenge, particularly in dynamic and multimodal contexts where appearance, motion, and context must be jointly modelled. Developing efficient frameworks that combine the scalability of self-supervision with the interpretability of semantic alignment is essential for achieving generalisable, semantically grounded video understanding.

1.2.3 Modelling Long-Term Temporal Dependencies

Fig 1.3 illustrates the evolution of video understanding tasks from short, trimmed clips to long, untrimmed and multimodal scenarios, motivating the need for long-term temporal and multimodal modelling.

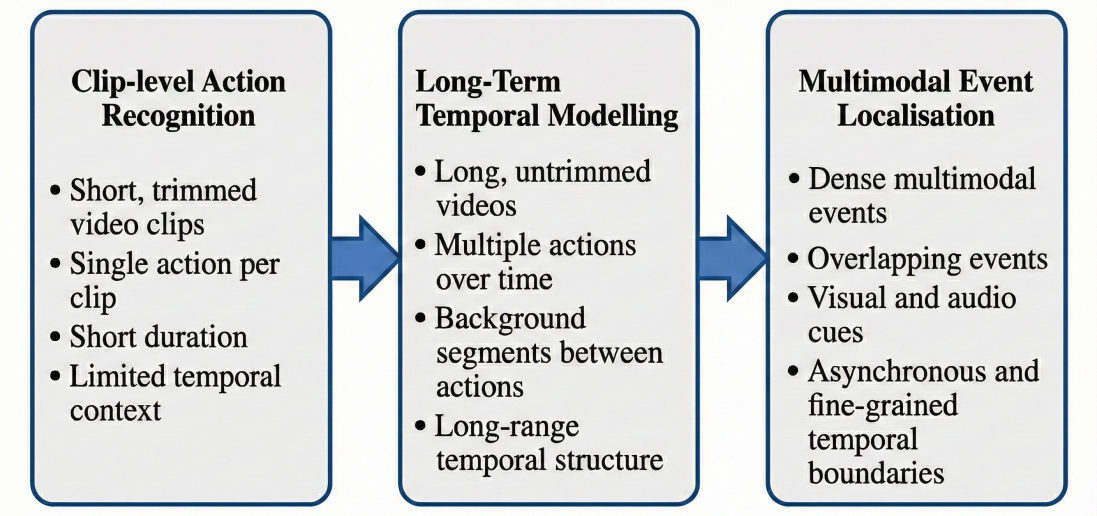


Figure 1.3: Evolution of video understanding tasks from clip-level action recognition to long-term temporal modelling and dense multimodal event localisation.

Understanding video content requires more than recognising visual patterns; it involves reasoning about how events evolve over time. Although long-term understanding may also depend on semantic abstraction beyond temporal structure, this section concentrates on the challenges associated with modelling extended temporal dependencies. Capturing temporal dependencies, particularly in long-term video understanding, remains a core challenge in the field [270,

104]. Real-world videos are inherently sequential, containing overlapping activities, scene transitions, and extended interactions that cannot be represented by short clips or isolated frames. While recurrent models attempted to capture such structure, they often struggled with long-term dependencies due to vanishing gradients and limited context windows [56, 306]. Transformer-based and hierarchical methods have improved temporal reasoning through self-attention mechanisms, yet maintaining efficiency and coherence across long sequences remains a challenge [17, 6, 271, 105].

The challenge becomes even more complex when events unfold across multiple time scales, where short, rapid motions occur alongside slow, evolving activities. Effective models must capture both fine-grained temporal dynamics and long-range dependencies to infer causal relationships and understand how actions and events develop over time. Temporal understanding also extends beyond visual cues to include cross-modal dependencies, such as how sound and contextual information evolve alongside visual changes. Achieving a balance between detailed temporal sensitivity and high-level abstraction is essential for coherent, causal, and robust video understanding in real-world scenarios.

1.2.4 Multimodal Fusion and Cross-Modal Dependencies

Videos are inherently multimodal, combining visual, auditory, and sometimes textual information that together form a complete and coherent understanding of events. Each modality contributes unique yet complementary cues: vision captures spatial and motion dynamics, audio conveys temporal continuity and environmental context, and language provides high-level semantics and intent. Integrating these heterogeneous sources remains a major challenge, as they often differ in temporal resolution, reliability, and relevance. For instance, audio cues may occur before or after corresponding visual events, or background noise may obscure meaningful relationships between modalities. An effective fusion strategy must therefore dynamically align and weigh these signals, ensuring that each contributes appropriately to the overall understanding.

Early approaches often relied on simple concatenation or fixed-stage integration, which limited the model's ability to capture complex dependencies and adapt to varying contexts [237]. More recent research has introduced attention-based and transformer architectures that enable flexible cross-modal interaction, allowing different modalities to interact and influence each other over

time [184, 197, 265]. In parallel, contrastive learning has emerged as a powerful technique for aligning visual and auditory representations, improving generalisation across tasks [230, 143, 5].

Despite these advances, effective multimodal fusion remains challenging, particularly for long, untrimmed videos where relevant cues are often sparse, asynchronous, or context-dependent. Models must learn to focus on informative moments, suppress noise, and establish meaningful temporal correspondences across modalities. Addressing these challenges requires adaptive frameworks that can learn cross-modal dependencies while maintaining temporal structure and computational efficiency. Such systems must not only integrate information from multiple sensory streams but also reason about how these signals evolve and interact over time. Achieving this integration is key to advancing real-world video understanding, enabling systems to perceive and interpret multimodal information in a cohesive, temporally consistent, and contextually meaningful way.

1.3 Problem Statements and Solutions

This thesis develops advanced learning frameworks designed to address fundamental challenges in video understanding, advancing toward comprehensive and multimodal analysis. Each framework addresses a key limitation of existing approaches while advancing a unified direction for learning from unlabelled data to achieve semantically grounded and multimodal reasoning. The research first explores how self-supervised learning can be guided to capture meaningful motion cues, enabling models to understand dynamic actions without manual labels. It further incorporates semantic guidance through language, enriching visual representations with conceptual meaning and improving interpretability. Finally, the study extends to multimodal learning by integrating audio and visual information to detect and temporally localise fine-grained, overlapping events in long, untrimmed videos, providing a more complete and context-aware understanding of complex real-world scenarios. This section introduces three main research problems along with their solutions, which will be discussed in detail in the following chapters.

1.3.1 Motion-Aware Representation Learning in Self-Supervised Video

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning visual representations without manual labels. However, existing SSL approaches in video understanding often fail to explicitly model motion, which is fundamental for recognising human actions. Standard masked autoencoders and contrastive methods typically focus on spatial regions or apply random masking, resulting in representations that emphasise static appearance rather than temporal change. However, motion plays a critical role in understanding actions, particularly in egocentric videos where subtle hand and object interactions convey crucial meaning. However, motion cues derived from optical flow are often corrupted by camera movement and background noise, leading to unstable learning signals and limiting SSL methods' ability to capture the fine-grained temporal dynamics required for robust action recognition.

Solution

To address this challenge, we propose MOFO, a motion-focused self-supervised framework that explicitly integrates motion saliency into video representation learning. MOFO introduces automatic motion area detection using unsupervised optical flow derivatives to reliably identify genuine motion regions while suppressing camera-induced noise. During pretraining, a motion-guided masking strategy prioritises masking within these dynamic regions, encouraging the model to focus its reconstruction objective on dynamic regions where actions occur. Furthermore, during finetuning, a multi-head cross-attention mechanism fuses embeddings from motion and non-motion areas, thereby strengthening motion-aware features for downstream recognition tasks. This design enables MOFO to learn semantically rich and temporally discriminative representations, achieving strong performance in self-supervised action recognition.

1.3.2 The Challenge of Learning Semantic Abstraction Without Labels

Achieving a high-level understanding of videos requires moving beyond raw pixels and motion patterns toward representations that capture meaningful semantic concepts such as activities, intentions, and interactions. However, in the absence of labelled supervision, most existing self-supervised methods remain limited to learning low-level visual patterns or instance-specific features. These models often allocate equal attention to all spatial and temporal regions, overlooking the crucial segments where the core actions occur. As a result, much of the learnt

representation focuses on background or transitional frames that contribute limited semantic meaning. Without explicit semantic guidance, models struggle to learn features that capture narrative structure, object-function relationships, or event relevance.

Recent progress in vision-language pretraining has shown that textual information can provide a powerful source of semantic supervision. Nonetheless, extending such methods effectively to video remains challenging, as many approaches depend on fully supervised video–text datasets or coarse global alignments that overlook fine-grained connections between dynamic motion cues and linguistic semantics. Consequently, the learnt representations often fail to capture the underlying meaning of events, limiting their ability to generalise across complex video understanding tasks.

Solution

We propose FILS, a self-supervised framework that learns semantically grounded video representations by aligning visual features with natural language. FILS replaces pixel-level reconstruction in masked autoencoding with a feature prediction objective implemented in a teacher–student framework, where the model predicts the features of masked video patches within a language-defined semantic space using textual descriptions. To incorporate linguistic semantics without manual labels, FILS leverages automatically generated video captions and introduces ActCLIP, a novel patch-wise video–text contrastive learning approach that focuses learning on the most informative parts of a scene, referred to as the action area. This patch-wise contrastive design enables the model to learn motion-aware, language-grounded representations that capture both visual dynamics and high-level conceptual meaning. Consequently, by combining motion-guided masking with language-based contrastive supervision, FILS produces semantically interpretable and transferable video representations that capture both motion and semantics, leading to strong performance on downstream tasks such as action recognition without requiring human annotations.

1.3.3 Dense Temporal Localisation of Multimodal Events in Untrimmed Videos

Real-world videos are rarely concise or neatly segmented. They often span long durations with multiple overlapping activities and diverse modalities, including visual scenes, speech, ambient sounds, and environmental noise. Unlike short, trimmed clips used in conventional recognition

or pretraining tasks, long untrimmed videos contain complex temporal structures in which events may overlap, occur asynchronously, or unfold at different time scales, making it difficult to define clear boundaries or align relevant signals across modalities. For instance, an audio cue may precede a visible action, or two unrelated activities may occur simultaneously in different spatial regions. This complexity makes it difficult to determine clear event boundaries or to associate relevant signals across time and modalities.

Audio provides complementary cues that enrich visual understanding by revealing information beyond what visual data alone can capture, indicating specific action timing, and distinguishing visually similar events. Despite the complementary nature of audio and visual information, most existing temporal action localisation methods focus solely on visual cues or rely on simple late-fusion techniques. These methods often assume perfect synchronisation between modalities and fail to capture the fine-grained temporal dependencies and complex cross-modal interactions needed for accurate event localisation. As a result, current systems struggle to distinguish overlapping actions, handle asynchronous events, and maintain precision in densely structured, untrimmed videos.

The key challenge, therefore, is to design a framework that can effectively model multiscale temporal structure and cross-modal dependencies within untrimmed videos. In this thesis, the challenge of long untrimmed videos is addressed specifically through dense localisation of overlapping events at multiple temporal scales, rather than through generic long-video modelling. Such a framework must be capable of dynamically aligning audio and visual features, capturing both global context and fine-grained temporal details, and accurately localising multiple concurrent events to provide a coherent, context-aware understanding of complex multimodal video data.

Solution

To overcome these limitations, we propose DEL, a unified multimodal framework for fine-grained audio-visual event localisation in untrimmed videos. DEL incorporates an adaptive masked attention mechanism to dynamically align audio and visual features, ensuring intra-modal consistency and cross-modal synchronisation. A dual contrastive learning objective further enhances discrimination by combining inter- and intra-sample losses with a score-based pair selection strategy to automatically identify informative positive and hard-negative examples.

In addition, a multiscale temporal fusion module aggregates hierarchical features, capturing both local temporal details and high-level contextual information. This design enables DEL to detect multiple overlapping or asynchronous audio–visual events with precise temporal boundaries. By effectively aligning heterogeneous modalities and modelling complex temporal dependencies, the framework can distinguish between overlapping sound sources and concurrent visual activities. For example, it can distinguish between overlapping sound sources such as background music and speech, or between concurrent visual activities such as human–object interactions. Through supervision with ground-truth event boundaries and class labels, DEL learns fine-grained temporal and cross-modal dependencies, achieving state-of-the-art performance (using the same pre-extracted feature inputs as previous methods) on standard benchmarks, demonstrating its strength in dense, multimodal reasoning for long, untrimmed videos, an aspect still insufficiently explored in current video understanding research.

Fig. 1.4 provides a conceptual summary of this thesis, illustrating how visual, auditory, and semantic information are jointly utilised to achieve comprehensive video understanding. A video inherently contains multiple information, including visual appearance, motion dynamics, and audio context, each contributing to how events are perceived and interpreted. The research presented in this thesis advances deep models' ability to capture and integrate these complementary cues. It begins with motion-focused self-supervised learning in **MOFO**, which enhances sensitivity to dynamic regions that are crucial for understanding actions. **FILS** extends this by grounding visual representations in language semantics, enabling richer contextual abstraction and semantic alignment. Finally, **DEL** provides a multimodal framework that aligns and localises complex audio-visual events in real-world videos. Collectively, these contributions move toward a model capable of perceiving, reasoning, and understanding video content in a coherent and human-like manner.

1.4 Thesis Outline

Building on the proposed solutions, this thesis is structured as follows.

Chapter 2: This chapter begins by reviewing relevant literature across the key areas of this research. This includes an overview of existing works in video understanding, covering the

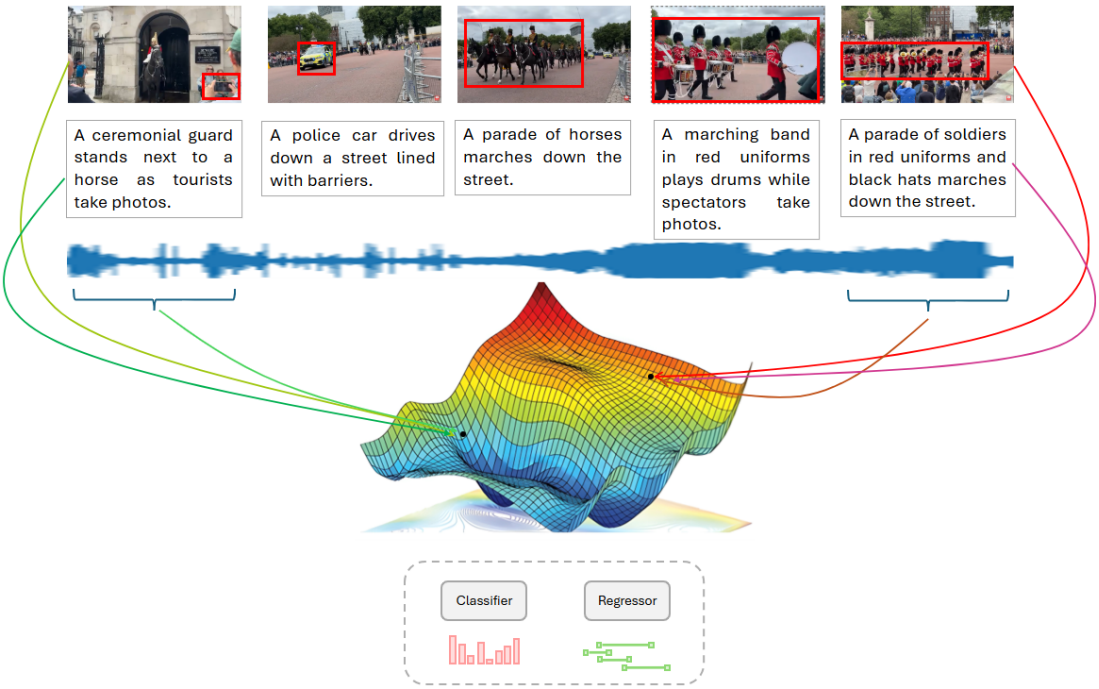


Figure 1.4: Conceptual overview of this thesis, depicting the integration of visual, auditory, and semantic information for holistic video understanding.

transition from handcrafted approaches to deep and transformer-based frameworks. We then discuss advances in self-supervised learning, including contrastive, reconstruction-based, and motion-aware methods, as well as progress in multimodal learning across vision–language and audio–visual domains. Finally, it highlights the challenges in egocentric video analysis and temporal event localisation, outlining the research gaps that motivate the proposed frameworks.

Chapter 3: In Chapter 3, a motion-focused self-supervised learning framework is introduced that explicitly incorporates motion cues into video representation learning. Motion regions are identified using spatial derivatives of optical flow, highlighting meaningful movement while reducing the influence of camera motion. These regions guide a motion-based masking strategy, encouraging the model to focus on dynamic, action-relevant areas during pretraining. During finetuning, a multi-head cross-attention mechanism fuses features from both inside and outside the detected motion area to enhance temporal reasoning and contextual understanding. Together, these components enable the model to learn motion-aware and semantically rich representations, thereby improving performance in action recognition tasks.

Chapter 4: This chapter introduces FILS, a self-supervised framework for video representation learning through feature prediction in a semantic language space. FILS constructs a unified embedding space that connects vision and language, enabling the model to learn semantically rich representations without manual annotations. The framework integrates two complementary objectives: masked autoencoding with feature prediction in the language space and ActCLIP, a motion-aware video–text contrastive strategy. ActCLIP focuses on patches within detected action regions and aligns them with automatically generated captions, ensuring that the learnt features capture both motion and semantic context. Unlike conventional masked autoencoders, which reconstruct pixels, FILS predicts features directly in the semantic space, promoting high-level contextual understanding. A teacher–student architecture with an exponential moving average further stabilises training, with the student predicting masked features while the teacher provides targets.

Chapter 5: Chapter 5 addresses the challenge of fine-grained audio–visual event localisation in untrimmed, real-world videos. The proposed multimodal framework integrates adaptive attention for cross-modal alignment, a dual contrastive learning objective for discriminative event modelling, and a multiscale path aggregation network for temporal fusion. The method enables precise detection and classification of overlapping and asynchronous audio–visual events, achieving superior performance on several benchmark datasets while maintaining a compact design with fewer parameters than comparable multimodal models.

Chapter 6: Chapter 6 concludes the research with a summary of the thesis contributions and the existing gaps in the proposed works. The chapter also outlines future directions for this research, focusing on developing more unified and end-to-end trainable frameworks, enhancing multimodal integration through language cues, and extending the proposed methods toward open-world generalisation and video understanding.

Chapter 2

Literature Review

This chapter presents a literature review, beginning with key historical contributions to the development of video understanding. After this overview, we move on to a more detailed discussion of recent studies that address the challenges and solutions introduced earlier.

2.1 Development of Video Understanding

Video understanding has evolved remarkably over the past decades. Early methods relied on handcrafted features such as SIFT [170], SURF [16], and HOG [42], along with temporal models like Hidden Markov Models (HMMs) [165] and traditional classifiers such as SVMs [220] and decision trees [304]. These approaches, though effective for low-level motion analysis, were limited in their ability to capture high-level semantic information.

The advent of deep learning led to early neural video models such as DeepVideo [117] and two-stream Convolutional Neural Networks (CNNs) [70, 223], which integrated appearance and motion cues. To model long-term dependencies, architectures like LSTM [306] and Temporal Segment Networks (TSN) [254] were introduced. Subsequently, 3D convolutional models, including C3D [239] and Inflated 3D ConvNets (I3D) [25], significantly advanced video classification benchmarks.

Transformers, proposed by Vaswani et al. [242], revolutionised neural network design by introducing an effective mechanism for modelling long-range dependencies in sequential data. The

key innovation of this architecture is the self-attention mechanism, which enables the model to dynamically identify and emphasise the most relevant elements within an input sequence. By allowing the entire sequence to be processed in parallel, transformers overcome the sequential processing limitations of recurrent networks [56] and the locality bias of convolutional networks [223, 233, 99]. As a result, they represent a significant advancement in efficiently capturing global contextual relationships across data. This innovation inspired Vision Transformers (ViTs) [59, 122], which have emerged as a powerful alternative to conventional convolutional networks. Their architecture builds upon the prominent Transformer encoder [50, 242] initially developed for natural language processing, where data are represented as sequences of vectors or tokens. Similarly, ViTs divide an image into a grid of non-overlapping patches, linearly project them into embeddings, and then process them through feed-forward and multi-head self-attention layers. This design enables ViTs to capture long-range dependencies and global contextual relationships beyond the reach of traditional CNNs. ViTs demonstrated strong performance in classification [313, 282, 134], object detection [34, 142], segmentation [41, 22, 12] and retrieval [75]. Inspired by ViT, several works extended transformers to video data. ViViT [6] and TimeSformer [17] introduced fully transformer-based video models that surpass traditional 3D CNNs. ViViT [6] defines the tubelet embedding tokenisation method and has inspired some other works to represent a video input by extracting non-overlapping, spatiotemporal tubes.

2.2 Representation Learning

Research in video understanding has also focused on improving how models learn from data. Early approaches primarily relied on fully supervised learning [117, 113, 223], which depends on large-scale annotated datasets. However, the cost and limited scalability of manual annotation have encouraged the exploration of more data-efficient paradigms, including unsupervised [173, 105], weakly supervised [130, 157], and self-supervised learning [229, 284, 238]. Among these, self-supervised learning has proven particularly effective, enabling models to learn rich spatiotemporal representations directly from raw video data and leading to substantial improvements in transferability and robustness across a wide range of downstream video understanding tasks.

2.2.1 Self-Supervised Learning for Video Understanding

Video understanding is a fundamental aspect of visual recognition, involving the analysis and interpretation of dynamic visual inputs. Deep learning approaches have significantly advanced these tasks; however, their success remains heavily dependent on large-scale, annotated datasets, which are both costly and time-consuming to create. Moreover, many real-world visual categories or contexts are not represented in common benchmarks, and their appearance may evolve over time, posing additional challenges for purely supervised approaches.

Self-supervised learning (SSL) has emerged as a promising direction to mitigate these limitations by reducing the reliance of machine learning models on extensive labelled data. Supervised approaches depend on high-quality annotations to guide model training, but producing such labels for large and diverse datasets is often infeasible. Consequently, the high cost of high-quality annotated data is a significant bottleneck in the training process. SSL leverages raw, unlabelled data to automatically generate supervisory signals. By designing pretext tasks, networks can learn meaningful and transferable representations without human-annotated labels. These pretext tasks produce pseudo-labels derived from data's intrinsic properties, enabling the model to capture structure, semantics, and temporal dependencies from raw inputs.

SSL first gained considerable popularity in natural language processing (NLP) [51], where models such as BERT demonstrated the power of learning from unlabelled text. The paradigm was later extended to computer vision [55, 33, 280], becoming one of the most influential approaches for representation learning. The approach is particularly relevant to video understanding, where obtaining dense, frame-level annotations is both expensive and labour-intensive. Self-supervised video learning aims to exploit the intrinsic spatiotemporal coherence of video data to learn representations that generalise across tasks. The central idea is to design auxiliary or pretext objectives that enable the network to learn robust intermediate representations from raw, unlabelled data without human supervision.

Earlier studies on video understanding focused on using RGB frames as input to learn action representations [251]. Convolutional Neural Networks (CNNs) were initially employed to extract frame-wise features, often aggregated via average pooling [117], which unfortunately discarded temporal order. To incorporate temporal information, frame-level CNN features were later processed with Long Short-Term Memory (LSTM) networks [56]. Two-stream networks [223]

further extended this approach by computing representations from both RGB frames and stacked optical flow frames. Subsequently, spatio-temporal 3D CNNs [239, 241, 69, 25] were introduced to model spatio-temporal patterns directly. The Persistence of Appearance (PoA) motion cue proposed in PAN [310] enabled networks to extract motion information directly from adjacent RGB frames, eliminating the need for explicit optical flow computation. Building on these foundations, recent advances in self-supervised video learning increasingly leverage temporal dynamics as a natural source of supervision. By designing pretext tasks that encourage temporal consistency, invariance, and predictiveness, these methods enable models to learn rich and generalisable video representations.

Representations learnt through these self-supervised objectives can then be finetuned for a wide range of downstream video understanding tasks, such as action recognition, temporal segmentation, and object detection [238, 141, 229, 100]. Numerous studies [71, 284, 249] emphasise the importance of modelling temporal dependencies, making SSL models sensitive to frame order, motion cues, and long-term context.

Representative paradigms include contrastive learning [307, 91, 297], self-distillation [22], and masked modelling [267, 92, 238, 82]. In contrastive approaches, models learn to distinguish between temporally or semantically related and unrelated clips; self-distillation methods promote consistency between augmented views of the same video; and masked modelling techniques, inspired by advances in natural language processing (NLP), train networks to reconstruct missing video regions, thereby encouraging the capture of contextual and temporal dependencies.

Ultimately, self-supervised video representation learning aims to enable models to extract semantic and temporal relationships directly from raw video data, mirroring the human ability to perceive and reason about structure and meaning without explicit supervision. By leveraging the intrinsic structure of unlabelled videos, SSL reduces annotation costs, enhances generalisation, and advances the development of scalable, data-efficient, and adaptable video understanding systems.

2.2.2 Reconstruction-Based Self-Supervision

Autoencoders [129] represent one of the earliest approaches to representation learning and can be interpreted as an early form of self-supervised learning through reconstruction objectives. A

variety of autoencoder architectures have been proposed in the literature, including the denoising autoencoder [129], the stacked denoising autoencoder [246], the contractive autoencoder [208], and the variational autoencoder [124].

Building on these developments, masked autoencoders [97] have recently emerged as a powerful yet conceptually simple and efficient variant and have proven an effective pretraining paradigm for Transformer models of text [50], images [97], and, more recently, videos [238]. Similar in spirit to denoising autoencoders, masked autoencoders corrupt the input image and aim to recover the original content. Specifically, the image is partitioned into non-overlapping patches, a random subset of which is masked, and the model is trained to reconstruct the missing patches. This approach encourages the learning of high-level semantic features while maintaining efficiency in large-scale visual representation learning.

Nowadays, encoder-decoder Transformer-based architectures are commonly used for self-supervised video representation learning. These architectures leverage the strengths of Transformer models, originally developed for natural language processing, and adapt them to process and comprehend video data. In the context of video representation learning, the encoder-decoder Transformer architecture typically consists of the following components:

1. **Encoder:** The encoder processes the input video data and generates a condensed representation of the video. Each video frame, or 3D tublet, is typically treated as a sequence of features to be input into the Transformer encoder. Multiple layers of self-attentional and feed-forward neural networks can be used in the encoder to capture the video's temporal dependencies, spatial relationships, and long-range dependencies.

2. **Decoder:** Based on the self-supervised task, the decoder generates a prediction using the encoder's learnt representation. The decoder must solve the surrogate task used for self-supervised learning. For instance, if the self-supervised objective is to anticipate the temporal order of shuffled frames, the decoder may correctly predict that order.

In transformer-based architecture, the self-attention mechanism powers both the encoder and decoder. Self-attention architectures are typically composed of a series of transformer blocks. Each transformer block consists of two sublayers: a feed-forward layer and a multi-head self-attention layer. An input is divided into patches, and attention evaluates each 3D input patch's usefulness

before drawing on it to produce the output. The Transformer's self-attention mechanism enables the model to focus on different parts of the video frames while considering their dependencies. Therefore, considering their relative importance, it draws from each input component to produce the output. The query(Q), key(K), and value(V) vectors are the three sets of calculated vectors in the transformer architecture. These are determined by multiplying the input by a linear transformation.

Recent generative self-supervised methods [97, 267, 281] have advanced the field. Masked auto-encoders (MAEs) [97, 83, 238] randomly mask input pixel patches and reconstruct them by minimising reconstruction errors in pixel space, showing competitive performance when finetuned. Other work reconstructs directly in the latent space [8, 10] or aims to predict contextualised latent representations that incorporate information from the entire input through masked prediction [10].

More recently, self-supervised learning has progressed beyond explicit reconstruction objectives. Joint-Embedding Predictive Architectures (JEPA) [8] and their video extensions, such as V-JEPA [14] and V-JEPA2 [9], are self-learning models that operate by predicting masked representations directly in latent space rather than reconstructing pixel-level inputs. JEPA was originally introduced as a general predictive framework applicable across modalities, emphasising the learning of semantically meaningful structure through latent prediction. V-JEPA adapts this principle to video by incorporating temporal prediction of spatiotemporal representations, enabling the modelling of motion and temporal coherence, while V-JEPA2 further extends this formulation through larger-scale training and longer-horizon temporal reasoning, improving transfer across diverse video understanding tasks. By focusing on predictable semantic structure instead of low-level visual detail, these approaches represent a shift from reconstruction-based self-supervision toward latent predictive learning.

2.2.3 Contrastive Learning

Contrastive learning has emerged as a powerful paradigm for representation learning, enabling models to learn discriminative features. The core principle involves bringing semantically similar (positive) samples closer in the embedding space while pushing dissimilar (negative) samples apart [188]. Given a data point called anchor, the method dynamically defines positive

and negative examples for the anchor during training. This objective encourages the model to learn robust, context-aware representations that generalise effectively across downstream tasks. Contrastive learning methods depend on a sufficient number of negative samples to learn high-quality representations. Negative samples play a critical role in enhancing the diversity of the training data, enabling the model to capture representative features of positive samples while learning to distinguish them from negatives. Furthermore, incorporating negative samples encourages the model to learn more discriminative and generalisable representations, helping to reduce overfitting and mitigate potential data biases. Conceptually, negative samples serve as perturbations within the training process, compelling the network to develop more robust and discriminative feature representations that capture the underlying complexity and variability of the data distribution. Nevertheless, identifying a sufficiently large and meaningful set of negative examples remains a significant challenge.

The Noise Contrastive Estimation (NCE) [93] and InfoNCE loss functions [188] are among the most widely used objectives in contrastive learning, forming the foundation for many modern frameworks. These losses optimise the similarity between positive pairs while contrasting them against a set of negatives, effectively transforming unsupervised learning into a classification-like problem. Variants such as symmetric addition and averaging have been proposed to improve training stability and convergence. Other formulations, including mean squared error (MSE) and cross-entropy, are also occasionally employed depending on the specific architecture and task requirements.

Early frameworks such as SimCLR [32] and MoCo [98] demonstrated the effectiveness of contrastive learning in the visual domain by leveraging large numbers of augmented image pairs and employing the InfoNCE loss [188] to optimise feature similarity. Fig. 2.1 illustrates the SimCLR framework, where two augmented views of the same image are generated and passed through a shared encoder and projection head. The model is trained to maximise agreement between these paired representations using a contrastive loss, and the learnt encoder is then used for downstream tasks. SimCLR introduced a simple yet scalable end-to-end architecture that relies on heavy data augmentation, while MoCo introduced a momentum encoder to maintain a dynamic memory bank of negative examples, improving consistency and efficiency during training. These methods laid the foundation for modern self-supervised learning, significantly advancing the learning of image and video representations.

Recent work has begun to explore score-based mechanisms within contrastive learning frameworks. For example, ScoreCL [123] introduces a score-matching function to estimate the strength of data augmentations and adaptively weight contrastive pairs in image-level representation learning. In this setting, the score reflects the discrepancy between augmented views and is used to modulate each pair's contribution during training, rather than encoding semantic relevance or temporal structure.

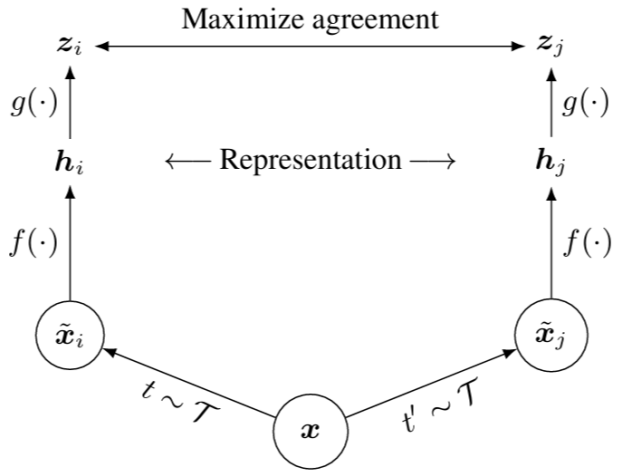


Figure 2.1: SimCLR framework [32] for contrastive visual representation learning.

In the multimodal domain, contrastive learning has become a foundational concept for cross-modal representation alignment, including vision–language, audio–language, and audio-visual settings [200, 36, 35, 143, 230, 57]. Models such as CLIP [200] and ALIGN [114] align visual and textual representations in a shared embedding space, enabling zero-shot transfer across a wide range of vision–language tasks. By jointly training on large-scale image–text pairs, these frameworks learn modality-invariant semantics that bridge the gap between natural language and perception. In the video domain, contrastive learning has progressed from capturing basic spatiotemporal cues to building large-scale multimodal foundation models. CVRL [198] established the basis for spatiotemporal contrastive learning by combining temporally consistent spatial augmentations with temporal sampling, ensuring that representations capture both motion and appearance. CLIP-VIP [289] adapted image–text pretrained models such as CLIP to the video domain by introducing a proxy-guided attention for temporal understanding and jointly training on video–subtitle and frame–caption pairs to improve video-text alignment. InternVideo [266] unified generative and contrastive learning based on the cross-model

learning between masked video autoencoding and video–text contrastive training, improving generalisation and multimodal understanding. LaViLa [322] further advanced this direction by leveraging large language models to generate dense video narrations, enriching temporal and semantic supervision for video–text alignment. More recently, GridCLIP [150] applies contrastive learning to open-vocabulary object detection through a dual-alignment strategy that links grid-level and image-level features. By aligning local spatial representations with CLIP's text encoder and distilling global semantics from its image encoder, GridCLIP achieves efficient multimodal learning with strong generalisation to unseen categories. it demonstrates how contrastive learning can evolve beyond representation learning to structured tasks such as detection, where local alignments strengthen global multimodal coherence. Contrastive learning provides a powerful foundation for both unimodal and multimodal representation learning. Its integration into multimodal systems has driven significant progress in vision–language understanding, audio–visual reasoning, and video-based event localisation, key research directions further explored in this thesis. In particular, in chapter 4, FILS applies contrastive learning only to regions that are semantically relevant to the ongoing action, aligning motion-relevant video features with their corresponding textual descriptions in a shared semantic language space. While recent methods such as CAV-MAE [5] improve audio–visual alignment through fine-grained contrastive objectives, they still lack effective mechanisms for temporal discrimination and adaptive hard-negative selection, limitations that are addressed in Chapter 5 through a dual and score-based contrastive framework.

2.3 Motion-Aware Video Understanding

Motion cues [1, 248, 138] have been recognised as essential for video understanding in the past few years. Optical flow serves as a prominent motion representation, effectively capturing the motion trends across consecutive frames.

Optical flow, first introduced by Lucas and Kanade [172], has become a foundational technique for motion estimation and feature extraction in computer vision. Their method assumes that small regions within an image exhibit coherent motion and formulates an optimisation problem to estimate pixel-wise motion vectors between consecutive frames. Subsequent developments, such as the Horn–Schunck approach [103] and Farnebäck's dense optical flow algorithm [65],

significantly improved the accuracy, density, and robustness of motion estimation. When integrated with predictive neural architectures, these optical flow methods have been effectively applied to downstream tasks, including object tracking, action recognition, and scene understanding. Building upon these advances, most works use optical flow, a motion representation component in many video recognition techniques, to obtain the statistical motion labels required for their work [292], separating the background from the main objects in optical flow frames. In another work [135], a multi-task motion-guided video salient object detection network is proposed, consisting of two sub-networks. One sub-network is used to detect salient objects in still images, and the other is used to detect motion saliency in optical flow images. Despite their effectiveness, traditional optical flow–based methods remain constrained by the assumption of locally uniform motion, which can result in degraded performance in scenarios involving rapid movements, large displacements, occlusions, or variations in illumination and noise.

Most motion descriptors use absolute motions and thus only work well when the camera and background are relatively static, such as Fleet & Jepson's phase-based features [73] and Viola et al.'s generalised wavelet features [247]. Therefore, the critical problem is identifying characteristics that accurately capture the motion of hands or objects while being impervious to the camera and backdrop motion.

2.3.1 Motion-Focused Self-Supervised Learning

Many video SSL studies have focused on enhancing the motion sensitivity of learnt representations. Several studies [78, 94, 107, 186, 279] employ optical flow to capture motion dynamics, while others [250, 54] achieve this by removing static or background components.

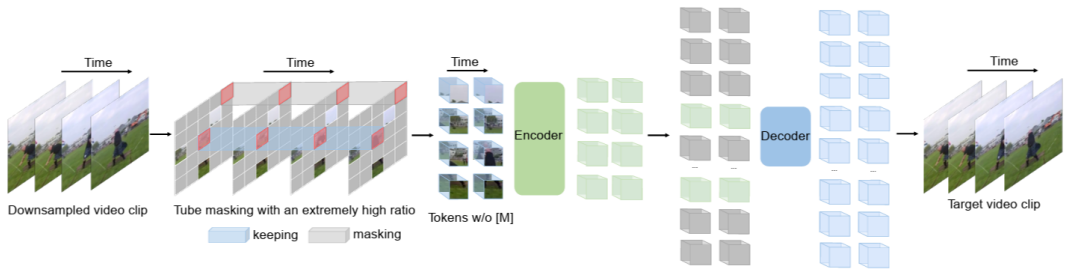


Figure 2.2: The VideoMAE framework [238] learns video representations by reconstructing highly masked spatiotemporal tubes using an asymmetric encoder–decoder architecture.

VideoMAE [238] introduced a simple yet effective masked autoencoding framework for self-supervised video representation learning. It learns spatiotemporal features by randomly masking a large portion of video tubelets and reconstructing the missing regions through an asymmetric encoder–decoder architecture. Fig. 2.2 illustrates the overall structure of the VideoMAE framework. Even though this model enables strong spatiotemporal reasoning over content, the encoder backbone could be more effective in capturing motion representations. Recent studies have further extended VideoMAE [238] by refining its architecture and training strategies to enhance motion understanding [63, 107, 231, 293]. For instance, MGMAE [107] and MGM [63] improve masking strategies by selectively masking motion regions identified using optical flow or motion vectors. MotionMAE [294] incorporates frame difference as an auxiliary reconstruction target, while MME [231] leverages optical flow to extract dense object trajectories and predicts both these trajectories and their corresponding Histogram of Oriented Gradients (HOG) features. Relying only on optical flow to capture the motion is not a robust solution, as it is heavily affected by camera motion and other external perturbations. To mitigate this problem, Wang et al. [248] presented a self-supervised spatiotemporal video representation by predicting a set of statistical labels derived from motion and appearance statistics by extracting optical flow across each frame and two motion boundaries, which are obtained by computing gradients separately on the horizontal and vertical components of the optical flow. Similarly, Han et al. [94] integrated optical flow with spatial features through a co-training scheme to enhance motion consistency learning. Beyond optical flow, residual frames have also proven effective in capturing motion-related features. For instance, Bi et al. [18] leveraged the consistency between residual motion paths and RGB visual streams to improve video representation learning.

The key contribution of our first project lies in explicitly integrating motion information into the masked autoencoding framework across both phases of self-supervised learning during pretext training and the subsequent finetuning stage. Additionally, we introduce an automatic motion detection mechanism capable of identifying salient objects and motion in the video without the overhead and limitation of a pretrained and annotated object detector.

Recent advances in multimodal learning have emphasised the need to adapt pretrained models to understand not only static visual semantics but also the dynamics of motion. Several studies have introduced finetuning strategies specifically designed to embed motion sensitivity into vision-language and video models. MotIF [111] finetunes a vision-language model using trajectory-

based visual representations, where robot motion paths are overlaid onto the final image frame. This motion-informed finetuning enables the model to associate visual trajectories with task and motion instructions, thereby learning trajectory-level reasoning from both human and robot demonstrations. MoCLIP [179] enhances CLIP's temporal understanding by finetuning its text encoder through contrastive learning with motion-sequence embeddings, combined with a feature distillation (tethering) loss to preserve CLIP's semantic knowledge while aligning its latent space with motion dynamics. DMGAL [90] introduces self and cross motion-guided attention modules during finetuning, enabling the model to identify and correlate motion-related features both within individual videos and across video sets. These motion-guided modules are trained through either full finetuning or adapter-based updates to efficiently capture spatio-temporal dependencies. These approaches demonstrate that targeted finetuning whether via trajectory visualisation, contrastive motion-text alignment, or motion-guided attention, can effectively endow pretrained models with a deeper understanding of motion dynamics.

2.4 Multimodal Video Understanding

Multimodal data refers to the integration of diverse sources of information such as images, text, and audio, that can be jointly processed by machine learning models to enhance performance across a wide range of tasks [15, 147]. By integrating these complementary modalities, multimodal fusion enables models to leverage the distinctive strengths of each source while compensating for the limitations or missing cues inherent in single-modality data. This synergy between modalities enriches representation learning, leading to more robust, flexible, and generalisable models across diverse application domains.

2.4.1 Vision-Language Representation Learning

In multimodal video-language learning and also language-guided video comprehension, incorporating language alongside videos [132, 74, 317, 229, 234, 322, 7] has introduced many intriguing challenges. Numerous attempts have been undertaken to merge computer vision and language, using the combined knowledge for various multimodal applications. Early work explored loss functions and architectures to grasp semantic vision-language alignments [256, 323]. Even before

deep learning became popular, early research investigated the process of learning visual representations from image captions [199]. Vision-language pretraining [171, 229, 228, 200] has become prevalent for learning transferable multimodal representations to enhance video-text tasks such as captioning [303], visual question answering [193], referring segmentation [299, 298] and others. Before masked autoencoders, contrastive learning jointly learnt vision and language representations. A notable example is Contrastive Language-Image Pretraining (CLIP) [200], which aligns image and text embeddings via a contrastive loss, achieving results comparable to supervised methods. It used modality-specific encoders that projected to a shared embedding space, with image-text pairs as targets. Work like [100] combined CLIP's language guidance with self-supervised contrastive learning to produce semantically aligned pixel embeddings, boosting finetuning and demonstrating language's generality in supervising vision. Recent works [57, 296, 140] explore incorporating masking into such pretraining, with strong results when finetuned on extensive labelled data. The standard vision-language pretraining pipeline, which involves initial pretraining followed by finetuning, is designed to develop a general multimodal feature representation suitable for multiple downstream tasks [267, 238, 10, 8].

The use of semantic embedding spaces has also been widely explored in zero-shot learning (ZSL), where the goal is to recognise unseen categories by transferring knowledge through class-level semantic representations such as attributes, word embeddings, or textual descriptions [261, 58]. Traditional ZSL methods learn mappings between visual and semantic spaces either via discriminative visual-to-semantic embeddings or through generative semantic-to-visual models that synthesise features for unseen classes [277, 245]. While effective for recognition transfer, these approaches fundamentally rely on externally defined class semantics and labelled class prototypes, which limits their applicability in fully self-supervised learning settings. In contrast, recent vision–language pretraining frameworks and the approach adopted in this thesis use semantic spaces not for unseen-class classification, but as a structure for representation learning. Rather than transferring knowledge across classes, semantic alignment is leveraged to structure visual representations and inject high-level meaning during self-supervised learning, enabling scalable representation learning without requiring explicit class annotations.

Related ideas from compositional modelling [79, 125] aim to represent visual concepts as compositions of reusable parts; however, such approaches typically rely on explicit structural

assumptions, whereas this thesis focuses on learning implicit structure through self-supervision and multimodal alignment.

While semantic alignment between vision and language provides a strong foundation for representation learning, many real-world video understanding problems require integrating additional modalities, particularly audio, which carries complementary temporal and contextual cues.

2.4.2 Vision-Audio Representation Learning

We live in a world surrounded by continuous streams of multimodal information, where visual and auditory cues jointly shape our perception and understanding of complex events. In recent years, there has been growing interest in audio–visual scene understanding, motivated by the need to interpret real-world environments where auditory and visual signals interact in complex and temporally overlapping interactions [268]. Existing research has primarily focused on tasks such as audio–visual action recognition [29, 278], audio–visual segmentation [77], and so forth, which typically operate on short video clips with limited temporal scope. More recently, fine-grained temporal event localisation has emerged as an important direction within the audio–visual learning community, aiming to recognise and temporally segment events across long, untrimmed videos [104]. Although several approaches incorporate both modalities [11, 110], they often treat one modality as auxiliary to the other, thereby neglecting the intricate cross-modal dependencies and long-range temporal relationships that characterise natural audio–visual interactions. This limitation hinders their ability to capture the richness and continuity of real-world multimodal dynamics.

Long-form audio–visual videos provide a natural and comprehensive record of such real-world dynamics, making them a crucial resource for advancing multimodal representation learning and scene understanding. These videos often span several minutes and capture multiple interrelated events that evolve or overlap over time, collectively contributing to the broader semantic narrative of the video. They inherently combine audio and visual modalities, which, while often asynchronous, provide complementary perspectives that enhance event comprehension and temporal reasoning. Effectively modelling the interplay between these heterogeneous and temporally misaligned signals remains a central challenge in developing robust audio–visual learning systems capable of understanding complex real-world scenarios. In Chapter 5, we address these

challenges by introducing a framework that learns to align and fuse audio–visual information, enabling coherent modelling of complex temporal dynamics and cross-modal interactions in long-form video.

2.4.3 Fusion Strategies

Multimodal fusion aims to integrate complementary information from different modalities, such as audio, visual, and textual signals, to build a unified understanding of complex video content. Deep learning–based fusion methods are generally categorised into early fusion and late fusion approaches, each offering distinct advantages and trade-offs depending on the task and modality characteristics.

Early fusion, also referred to as feature-level fusion, combines unimodal features at the input or intermediate representation stage, before the learning of semantic concepts. In this paradigm, features extracted from individual modalities are concatenated or projected into a shared feature space to form a joint multimodal representation. The integrated features are then processed through a single learning model, allowing it to capture interdependencies across modalities directly. Early fusion provides a truly multimodal representation by integrating at the feature level, enabling the model to jointly learn modality correlations from the outset. This approach has shown strong performance in tasks where cross-modal relationships are tightly coupled, such as emotion recognition from videos, where facial expressions, speech tone, and linguistic cues are fused early to infer affective states [13]. Similar strategies have been employed in action recognition by combining spatial and temporal streams [223], and in multimodal video retrieval, where fused audio–text–visual embeddings improve retrieval accuracy [180].

However, early fusion poses notable challenges when modalities differ substantially in structure, scale, or temporal dynamics. Early fusion methods also rely on temporally synchronised and complete multimodal inputs, which makes them sensitive to misalignment, missing data, and noise commonly found in real-world multimedia environments.

Aligning heterogeneous features in a common space often requires careful normalisation or transformation, and the fusion process may obscure modality-specific nuances. Furthermore, feature-level integration typically demands synchronised and complete data across modalities, limiting its robustness in real-world scenarios with missing or noisy signals.

Late fusion, also known as decision-level fusion, combines information after independent learning on each modality. Instead of merging raw or intermediate features, late fusion aggregates the outputs such as class probabilities or confidence scores, from modality-specific models. The final decision is made by integrating these predictions through operations such as weighted averaging, voting, or trainable fusion networks.

This approach leverages the individual strength of each modality, allowing specialised models to capture unique modality-specific information before combining their insights. For example, in video sentiment analysis, separate models analyse visual content, speech tone, and textual transcripts, with their outputs fused at the decision level to predict sentiment [194]. Late fusion has also enhanced large-scale video classification by merging spatial and temporal CNN outputs [117], improved video-based action recognition by integrating predictions across modalities [272], and aided cross-modal video retrieval by integrating textual and visual information during the final decision stage [159].

While late fusion offers a flexible, modular approach to combining multiple modalities, it also has inherent limitations. Each modality must be trained separately under supervision, which increases computational overhead. Because fusion occurs after individual processing, fine-grained temporal and semantic relationships between modalities may be weakened. Additionally, decision-level integration often requires an extra learning stage to combine unimodal outputs effectively, further increasing overall training complexity.

In summary, early fusion enables deep feature-level interaction across modalities but struggles with heterogeneous data alignment, whereas late fusion maintains modality-specific expressiveness at the cost of reduced cross-modal coherence.

Attention-based and multiscale fusion mechanisms have become fundamental in multimodal learning, offering an effective means of capturing complex dependencies between audio and visual modalities. Conventional fusion approaches typically concatenated audio and visual features before applying temporal modelling. Lin et al. [156] employed a bidirectional long short-term memory (Bi-LSTM) network to fuse sequentially encoded features, improving temporal coherence. Ramaswamy et al. [203] extended this idea by introducing attention-based modules to capture both local and global interactions between modalities. Xuan et al. [287] addressed temporal inconsistency using adaptive attention, while Duan et al. [60] proposed a joint co-

attention mechanism to model inter- and intra-modal relations through recursively stacked attention blocks. Lin et al. [158] further introduced an audiovisual transformer that unifies feature encoding and fusion via cross-modality co-attention, effectively associating audio cues with intra- and inter-frame visual information for event localisation. However, such dense attention designs often incur high computational costs.

Despite these advances, most existing methods [157, 285, 287] model interactions globally, overlooking variations in event duration and temporal dynamics common in real-world settings. Modelling semantics across multiple temporal scales may thus be essential for accurate multimodal understanding. Although several works explored multiscale structures [115, 301], they typically process scales independently and perform late fusion, limiting cross-scale interaction and requiring extensive manual tuning. More recent research, such as [316], advanced multiscale temporal fusion by generating features at multiple temporal resolutions and introducing a searchable fusion module to learn optimal cross-modal interactions, thereby improving the modelling of audio–visual asynchrony and event localisation accuracy. [301] proposed an attentive feature pyramid designed to capture and integrate multi-level temporal features for enhanced temporal reasoning. Jiang et al. [115] developed a hierarchical context modelling network that extracts contextual information across different temporal scales and employs a guiding network to learn discriminative modality-specific representations. However, their framework models each scale independently and performs late score fusion, which limits cross-scale and cross-modal interactions. Moreover, attention-based frameworks, though capable of highlighting salient features and mitigating modality inconsistencies, often fail to capture fine-grained events that are spatially subtle, temporally misaligned, and easily obscured by concurrent actions, which is a critical issue for realistic multimodal understanding. In Chapter 5, we address these limitations through our proposed framework, which introduces a Path Aggregation Network (PAN) to enable unified multiscale fusion and fine-grained cross-modal alignment. The PAN module aggregates features across multiple temporal resolutions, enhancing temporal granularity while preserving cross-modal coherence. By enabling dense temporal reasoning, DEL effectively captures subtle variations in event duration and mitigates audio–visual asynchrony, leading to more precise modelling of fine-grained and overlapping events in long-form videos.

2.5 Benchmark Datasets for Video Understanding

Benchmark datasets play a central role in video understanding research, as they define the visual complexity, temporal structure, and semantic scope that learning algorithms must address. As the field has evolved, benchmark design has progressed from short, trimmed clips toward long, untrimmed, and increasingly multimodal videos that better reflect real-world conditions. This evolution closely parallels advances in video models, which now aim to reason about motion, semantics, and cross-modal interactions over extended temporal horizons.

Early action recognition benchmarks such as UCF-101 [225] and HMDB-51 [127] were instrumental in establishing deep learning approaches for video classification. However, these datasets are limited in scale and temporal complexity, as actions are short, well-centred, and visually distinctive. Large-scale datasets such as Kinetics-400/600/700 [118, 23, 24] significantly expanded action diversity and dataset size, enabling the learning of strong generic video representations. Nevertheless, these datasets largely consist of trimmed clips and often allow models to rely on static appearance or scene context, reducing their effectiveness for evaluating temporal reasoning and fine-grained motion understanding.

To explicitly emphasise temporal dynamics, motion-centric benchmarks were introduced. Something-Something V2 [87] is a representative example, in which actions are visually similar and can only be distinguished through subtle motion cues and object interactions. This design discourages appearance-based shortcuts and has become a standard benchmark for evaluating motion-aware representations and temporal reasoning in video understanding models.

A further shift in benchmark design is the move toward egocentric video understanding, which captures activities from a first-person perspective. EPIC-KITCHENS-100 exemplifies this direction, providing long, untrimmed recordings of daily activities with dense annotations for verbs, nouns, and actions [44, 47]. Egocentric videos introduce challenges such as strong camera motion, frequent occlusions, motion blur, and fine-grained hand–object interactions, making them particularly suitable for evaluating models that explicitly capture motion and context under realistic conditions. More recently, Ego4D [88] has substantially expanded this domain by releasing thousands of hours of egocentric video across diverse environments and activities. Ego4D supports a wide range of tasks, including action recognition, long-term temporal reasoning, episodic memory, and audio–visual grounding, and places strong emphasis on real-world

variability and long-horizon understanding. Additional egocentric datasets such as Charades-Ego [222], EGTEA Gaze+ [144], MECCANO [202], and Assembly101 [213] further enrich this landscape by focusing on cross-view learning, gaze supervision, and procedural activities.

Beyond action recognition, temporal action localisation benchmarks extend the task to untrimmed videos, requiring precise prediction of action boundaries. THUMOS14 [112] remains a challenging benchmark due to its short action instances embedded within long background sequences. ActivityNet v1.3 [21] scales this setting to a larger and more diverse set of actions with substantial variation in action duration, while HACS [319] provides dense annotations to support more realistic evaluation of long-range temporal reasoning.

Recent work has also highlighted the importance of multimodal benchmarks that integrate audio and visual information. Early audio–visual datasets such as AVE [237] focus on short, trimmed clips containing a single dominant event, limiting their realism. In contrast, UnAV-100 [81] provides long, untrimmed videos with densely annotated audio–visual events that may overlap or occur asynchronously across modalities. These characteristics make it a challenging benchmark for evaluating fine-grained temporal localisation and cross-modal alignment. Egocentric datasets such as EPIC-KITCHENS-100 and Ego4D further support audio–visual learning by capturing rich auditory cues that naturally complement visual motion and interaction patterns.

The datasets considered in this thesis reflect the broader evolution of video understanding benchmarks toward greater realism, temporal complexity, and multimodal richness. Emphasis is placed on datasets that require explicit modelling of motion and temporal dynamics rather than reliance on static appearance cues, as well as benchmarks that capture realistic interactions and long-range temporal structure. Long, untrimmed videos with dense temporal annotations and, where applicable, multimodal signals provide a meaningful setting for evaluating robust video understanding under real-world conditions.

2.6 Application-Specific Review

In addition to the general progress in video understanding, several application areas pose distinct challenges that shape how these methods are developed and applied. The following sections

discuss key examples, including egocentric video analysis, action recognition, and temporal event localisation.

2.6.1 Characteristics and Challenges in Egocentric Video Analysis

Egocentric visual data, captured from wearable cameras or first-person perspectives, provides a uniquely human-centred view of interactions with the surrounding world. This perspective is inherently rich in multimodal cues such as gaze, motion, and hand-object interactions, offering an immersive understanding of human behaviour and cognition. However, these benefits come with significant challenges. Egocentric data is highly dynamic, characterised by rapid viewpoint shifts, frequent occlusions, and intense motion blur caused by natural head or body movements [139]. Such conditions complicate visual recognition, tracking, and segmentation tasks. Fig. 2.3 illustrates an egocentric video sequence of a person closing a cupboard, highlighting typical challenges of first-person vision such as motion blur, hand occlusion, and abrupt viewpoint shifts, all of which complicate consistent visual understanding and temporal reasoning.

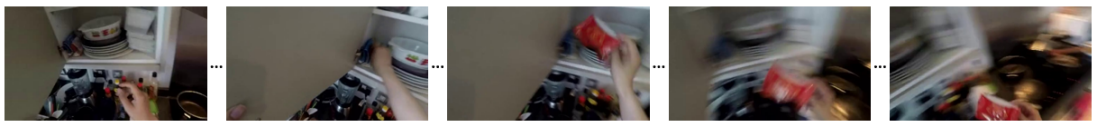


Figure 2.3: Example of an egocentric video sequence from the EPIC-KITCHENS-100 dataset showing a person closing a cupboard. The sequence highlights typical challenges in egocentric vision, including hand occlusion, rapid viewpoint changes, and motion blur, which complicate object detection and action recognition.

Moreover, the semantic structure of egocentric videos spans multiple levels of granularity, ranging from coarse procedural activities to fine-grained hand-object interactions, which makes modelling temporal dependencies and contextual consistency particularly challenging. Long-duration recordings introduce additional difficulties, as models must autonomously identify meaningful events within continuous streams of largely irrelevant content. Finally, issues such as personalised viewpoints, lighting variations, and inconsistent environments amplify domain variability, demanding more robust, data-efficient, and cross-view learning approaches.

A number of existing methods leverage object detection to improve egocentric video recognition [262, 270, 178], among which [270] also incorporates temporal contexts to help understand

the ongoing action. These approaches may have limited use in real-world systems, as they require time-consuming, labour-intensive item-detection annotations and are computationally expensive. In contrast, our proposed framework in Chapter 3 does not depend on costly object detectors. Recently, Shan et al.[216] developed a hand-object detector to locate the active object. When the detector is well-trained, it can be deployed on the target dataset; however, running the detector on high-resolution frames still costs far more than using our method.

While existing egocentric datasets have provided valuable insights into visual perception and human–object interaction, their limited scale and focus on specific environments restrict their usefulness for training robust deep learning models. To overcome these limitations, the EPIC-KITCHENS dataset [44], later extended to EPIC-KITCHENS-100 [47], was introduced as a large-scale egocentric video benchmark. It has since become the standard dataset for diverse egocentric vision tasks, including action recognition [84, 120, 290], action localisation [311, 219, 217], and action anticipation [209, 53]. Other datasets, such as MECCANO [202, 201] and Assembly101 [213], focus on procedural or industrial activities, capturing fine-grained assembly actions with synchronised multimodal data. The EGTEA Gaze+ dataset [144] focuses on egocentric cooking activities with synchronised gaze tracking, detailed action annotations, and hand masks. The Charades-Ego dataset [222] pairs first-person and third-person views of daily indoor activities, enabling cross-view understanding and egocentric action recognition. The large-scale Ego4D dataset [88] further expands the domain by covering thousands of hours of daily-life egocentric recordings across diverse real-world scenarios. The EPIC-KITCHENS dataset is utilised throughout this thesis (Chapters 3–5) to train and evaluate the proposed methods. Additionally, the Charades-Ego and EGTEA datasets are employed in Chapter 4 to further assess performance on egocentric action recognition.

2.6.2 Action Recognition in Egocentric and Complex Videos

Human action recognition is a longstanding challenge in computer vision, aiming to automatically identify and interpret human activities from visual data. Although it is simple for humans to recognise and categorise actions in video, automating this process is challenging. Human action recognition in video is of interest for applications such as automated surveillance and security [121, 128] detecting anomalies in a camera's field of view that has attracted attention

from vision researchers [243], elderly behaviour monitoring [211], human-computer interaction, content-based video retrieval [226], sports performance analysis [274] and video summarisation [215]. Activity analysis must be able to identify atomic movements such as ”walking”, ”bending”, and ”falling” on its own while monitoring the daily activities of elderly people, for instance [214]. Therefore, action recognition is a challenging problem with many potential applications.

2.6.3 Temporal Event Localisation in Untrimmed Videos

Temporal event localisation is the process of identifying the temporal boundaries of actions within a video sequence, a core task for advancing intelligent video understanding. This task extends beyond basic recognition, as it requires a more comprehensive understanding of actions in order to accurately determine their onset and offset times.

Early approaches to temporal action localisation relied on a range of hand-crafted techniques, including space-time features [86], spatiotemporal graphs [19], and hidden Markov models [236]. With the advancement of deep learning research [126], [224], the field has experienced a paradigm shift, leading to more robust and scalable solutions.

Single-Modality Temporal Localisation Tasks

Deep learning has significantly advanced temporal action localisation (TAL), enabling the detection and classification of actions in untrimmed videos. TAL methods are generally categorised into two-stage and single-stage approaches. Two-stage methods generate action proposals with confidence scores before refining and classifying them [61, 152, 153, 320, 163]. In contrast, single-stage approaches directly localise actions in a single pass, eliminating the need for proposal generation. Recent advances in single-stage temporal action localisation (TAL) have achieved greater accuracy and efficiency by integrating action proposal generation and classification into a unified, end-to-end process within a single forward pass.

These methods can be further classified into anchor-based [28, 168, 318, 258] and anchor-free approaches [148, 295, 183, 275, 276]. Anchor-based methods employ predefined multiscale temporal anchors to detect potential action segments, refining their boundaries via regression

techniques. Conversely, anchor-free methods predict action boundaries directly by modelling temporal dependencies or regressing distances to action start and end points.

Recent advances integrate graph neural networks (GNNs) [309, 295, 286] and transformers [255, 235, 27] to enhance temporal modelling. Transformers, in particular, have demonstrated superior performance by leveraging self-attention mechanisms to capture long-range dependencies. Inspired by progress in object detection [207, 38] and saliency detection [148], recent state-of-the-art approaches incorporate transformer-based feature pyramid networks (FPNs) [37, 269, 311, 218]. These networks effectively combine multiscale feature representation with classification and regression heads, improving accuracy and efficiency over conventional methods.

Video Temporal Grounding and Moment Retrieval

Video temporal grounding, also known as video moment retrieval, addresses the problem of localising a temporal segment in an untrimmed video that corresponds to a given natural language query. Unlike temporal action localisation, which detects and classifies events from a predefined label set, temporal grounding is driven by open-vocabulary textual descriptions and requires aligning visual content with linguistic semantics.

Early works such as MCN [3] and CTRL [76] formulated temporal grounding as a cross-modal matching problem between video clips and sentence embeddings. Subsequent methods improved temporal reasoning by introducing proposal-based frameworks [305], attention mechanisms [76], and multi-scale temporal modelling [312]. More recent approaches leverage transformer architectures to model long-range dependencies and fine-grained cross-modal interactions, including VSLNet [315], Moment-DETR [131], and SeqPAN [314].

With the rise of large-scale vision–language pretraining, grounding methods have increasingly adopted contrastive and alignment-based learning strategies. Models such as CLIP4Clip [177], UniVL [176], and X-CLIP [185] improve temporal grounding by leveraging shared embedding spaces between video and text. Recent works further explore fine-grained alignment through hierarchical reasoning and temporal interaction modelling, including HiT [162].

Although video temporal grounding and temporal action localisation both aim to localise events in time, they differ fundamentally in their supervision and alignment mechanisms. Temporal

grounding is query-driven and focuses on semantic alignment between visual content and natural language, whereas TAL is category-driven and emphasises temporal discrimination among visual events using predefined action labels. Consequently, grounding methods are primarily designed for cross-modal retrieval between vision and language, while TAL methods target dense event detection and boundary regression. Importantly, most grounding approaches overlook the role of audio, despite its strong temporal cues in real-world videos. This motivates the shift toward audio–visual event localisation, where localisation is guided by intrinsic interactions between sound and vision rather than by textual queries, enabling more robust handling of overlapping events, temporal asynchrony, and complex multimodal dynamics.

Multimodal Audio-Visual Event Localisation

Audio-visual event localisation aims to detect events that are simultaneously audible and visible in video content, thereby capturing their multimodal correspondence. Recent studies have shown that leveraging multiple modalities generally yields superior performance than relying on a single source of data, with particularly notable gains in temporal action localisation [11, 130, 101].

While joint audio-visual representation learning has been extensively studied in video retrieval and classification [120, 278, 2], its application in TAL is less explored, primarily due to the challenges of modality misalignment and the need for fine-grained temporal modelling. Most existing methods assume that the audio and visual events are perfectly aligned [237, 11, 300], limiting their effectiveness in real-world settings where events may be asynchronous or overlapping. As highlighted in our introduction, real-world scenarios often involve a complex interplay between modalities and concurrent events.

Conventional audio-visual TAL models employ two-stage late fusion strategies, where audio and visual modalities interact only at the final classification stage. This approach reduces their ability to model fine-grained temporal dependencies, making them less effective for complex event localisation. Some recent approaches incorporate cross-attention mechanisms [204, 288, 273, 161] but operate at fixed temporal scales and lack dynamic fusion control.

Furthermore, contrastive learning techniques [106, 265] often focus on instance-level alignment while neglecting intra-video relationships, such as temporal coherence and cross-event correlations, which are crucial for distinguishing similar events occurring at different times. This

limitation affects their ability to handle multiscale variations and modality-specific characteristics. In Chapter 5, we explicitly model cross-modal dependencies to address these challenges while preserving fine-grained temporal structure through an adaptive attention mechanism and a path aggregation network. Moreover, our dual contrastive learning strategy, incorporating inter-sample and intra-sample contrastive loss, refines feature discrimination within a single video, addressing the limitations of existing methods.

In summary, this chapter traced the evolution of video understanding from handcrafted features to self-supervised and multimodal learning. While existing research has made significant progress in representation learning, motion modelling, and cross-modal fusion, challenges remain in jointly capturing motion, semantics, and temporal dynamics. These gaps motivate the methods proposed in this thesis, which advance motion-focused self-supervision (Chapter 3), language-guided feature learning (Chapter 4), and dense multimodal event localisation (Chapter 5).

Chapter 3

MOFO: MOtion FOcused Self-Supervision for Video Understanding

Action recognition is an essential task in video understanding and has been extensively investigated in recent years [166, 267, 82]. This field has seen remarkable advancements with the advent of supervised deep learning methods, which have achieved significant success in leveraging labelled datasets for training [239, 68, 149]. However, the reliance on annotated data presents a major bottleneck. Annotating videos is labour-intensive and expensive, especially given the temporal complexity of actions and the vast number of frames involved. Due to the lack of labels, which must be manually collected, learning to recognise actions from a small number of labelled videos is a difficult task, as data collection will be expensive and challenging. It is especially inappropriate for long-tail distributions across scenes, such as kitchen activities, where rare actions are under-represented. Furthermore, getting annotations for videos is much more difficult due to the large number of frames and the temporal boundaries of when actions begin and end. Therefore, self-supervised learning (SSL) has gained attention due to the problems above and has demonstrated impressive results in learning visual representations across various domains. In SSL, a model is trained on data and labels, but in this case, the labels come from the data itself without expensive and limiting human annotation.

Supervised methods [264] have recognised the importance of motion for understanding actions because key objects often move in the scene. However, most SSL methods do not explicitly consider motion or use handcrafted features [62], thereby limiting their effectiveness. For SSL, masked autoencoder models [238] have been proposed to learn the underlying data distribution in a self-supervised manner, reconstructing spatiotemporal content without explicitly focusing on motion dynamics. Even though this model can perform spatiotemporal reasoning over content, the encoder backbone is ineffective in capturing motion representations (as shown later in fig. 3.3). Incorporating motion information is inherently challenging, since a major limitation of using optical flow for motion detection, particularly in egocentric video, is its sensitivity to camera movement. When the camera moves rapidly, static objects or background regions may appear to have high motion velocities, resulting in unstable and misleading optical flow estimates.

In this context, we propose detecting salient objects and motion in the video without the overhead and limitations of a pretrained and annotated object detector. This automatic approach for object detection is based on motion boundaries derived from optical flow. Using the motion boundaries instead of a direct optical flow output mitigates the challenge of camera motion and creates salient areas of movement or interest without a pretrained network. To fully exploit this phenomenon, in this work, we propose a motion-aware self-supervised approach, **MOFO** (MOtion FOcused Self-Supervision for Video Understanding), to perform self-supervised video action recognition. The key contribution is to explicitly incorporate motion information in both SSL phases of self-supervised pretext training without human annotations, and then in the finetuning stage. Given motion identification, we propose advancing motion understanding through the now-familiar self-supervised masking [238], where a large proportion of regions or 3D patches in video frames are masked based on motion cues as a self-supervision pretext task.

Since motion regions carry rich information about moving objects, actions, and interactions, MOFO prioritises these areas through its core innovation, a motion-guided masking strategy within a masked autoencoder framework. MOFO addresses the limitations of current SSL methods by automatically detecting motion regions and incorporating this information into both pretraining and finetuning stages. During the pretext task, a higher proportion of patches within the detected motion areas are masked, while the remaining regions are masked randomly. This motion-focused masking encourages the network to concentrate on motion-relevant fea-

tures, enhancing its ability to capture spatiotemporal structures relevant to action recognition. Additionally, during finetuning, MOFO further emphasises motion by employing multi-head cross-attention mechanisms that fuse embeddings from both within and outside the motion areas. This dual-stage focus on motion ensures that the learnt representations are not only robust but also contextually enriched.

In summary, MOFO's contributions are as follows:

- The automatic motion area detection using motion maps driven by optical flows, but invariant to camera motion.
- A motion-aware SSL approach, which focuses masking on the motion area in the video, using our proposed automatic motion detection algorithm.
- A motion-focused finetuning technique to further intensify the focus on the motion area for the action recognition task.
- The demonstration of our model's state-of-the-art performance for SSL action recognition on two large benchmark datasets, EPIC-KITCHENS-100 and Something-Something V2.

To evaluate the effectiveness of MOFO, we conducted extensive experiments on two challenging datasets: Something-Something V2 (SSV2) and EPIC-KITCHENS-100. These datasets involve egocentric videos with complex motions and limited labelled data, making them ideal benchmarks for testing SSL approaches. Our results demonstrate that MOFO significantly outperforms state-of-the-art SSL methods in terms of action classification accuracy. For instance, MOFO achieves a 4.7% improvement in Top-1 accuracy on SSV2 and notable gains across verb, noun, and action classifications on EPIC-KITCHENS-100 compared to leading baselines like VideoMAE [238] and OmniMAE [83].

3.1 Methodology

This section outlines our proposed SSL methodology, focusing representation learning on the motion area of a video for action recognition. Fig. 3.1 overviews our method, with three parts: First, our automatic motion area detection, with optical flow input to create a motion map

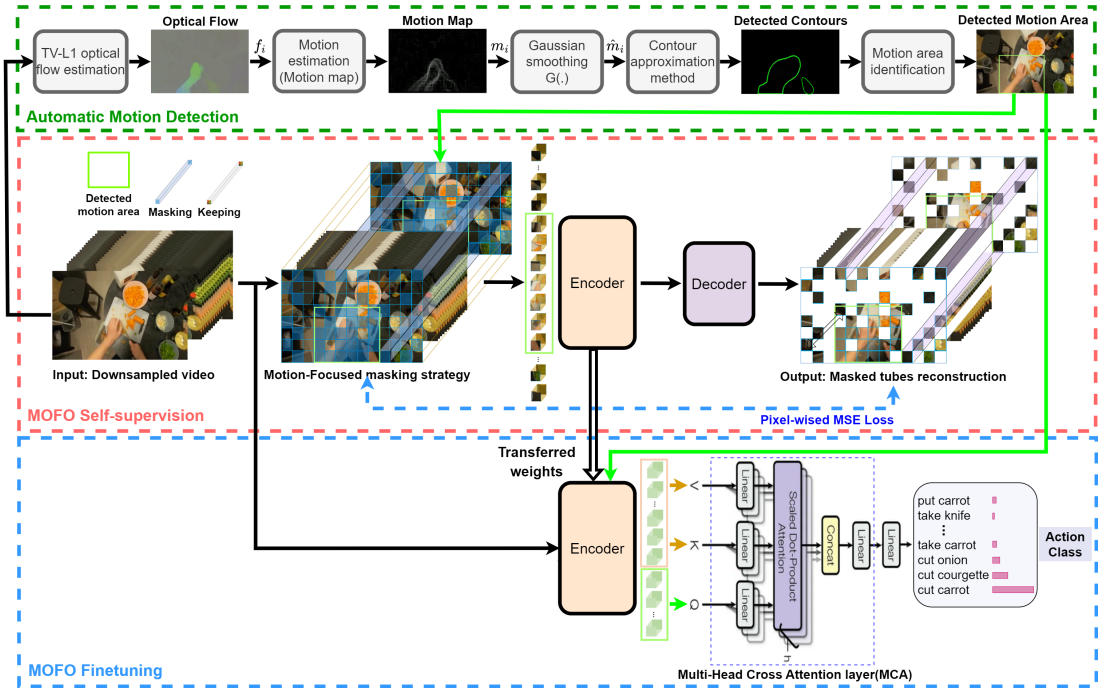


Figure 3.1: MOFO is a motion-focused self-supervised framework for action recognition.

to remove camera motion. Second, we propose our new strategy for the SSL pretext task, a reconstruction task that focuses more on masking 3D patches on the motion area in the video. Thirdly, the downstream task adaptation step emphasises motion further by integrating motion information during the training. Notably, the raw video is used only to localise motion regions; the model itself operates on a single encoder representation, which is later decomposed into motion-centric and contextual embeddings and fused via multi-head cross-attention for classification.

3.1.1 Automatic Motion Area Detection

To identify motion regions without relying on pretrained object detectors, we propose using classical computer vision cues derived from optical flow. However, optical flow vectors are sensitive to camera motion, causing static objects or background pixels to appear as if they have high motion velocities when the camera moves rapidly. To mitigate the problem above, we calculate the motion boundaries [43] and use these to define a motion map [138]. Therefore, given a video with T frames and a $H \times W$ dimension, we first extract the optical flow vectors representing $\{f_i \in \mathbb{R}^{H \times W}\}_{i=1}^{T}$ pixel-level motion between two consecutive frames in a video

using the TV-L1 algorithm [308] that offers increased robustness against illumination changes, occlusions and noise. Then, given the horizontal and vertical displacements of each pixel between the ith frame and the $(i+1)$th frame represented by the flow maps $u_i, v_i \in \mathbb{R}^{H \times W}$, any kind of local differential or difference of flow cancels out most of the effects of the camera rotation. The resulting motion map is defined as:

$$m_i = \sqrt{(\frac{\partial u_i}{\partial x})^2 + (\frac{\partial u_i}{\partial y})^2 + (\frac{\partial v_i}{\partial x})^2 + (\frac{\partial v_i}{\partial y})^2} \tag{3.1}$$

where every component denotes the corresponding x- and y-derivative differential flow frames contributing towards computing m_i, representing moving velocity in the i-th frame while ignoring the camera motion. As a result, $m_i \in \mathbb{R}^{H \times W}$ is less influenced by camera motion and considers the moving salients in the i-th frame. A low-pass Gaussian filter is used to smooth areas of the image with high-frequency components, further reducing the unwanted noise effect. The Gaussian smoothing operator computes an average of the surrounding pixels that are weighted according to the Gaussian distribution (G), which is as follows:

$$\begin{gathered} G_{x,y} = \frac{1}{2\pi\sigma^2} \exp^{(-\frac{x^2+y^2}{2\sigma^2})} \\ \hat{m}_i = G * m_i \end{gathered} \tag{3.2}$$

where $\hat{m}$ is the convolved motion map by the Gaussian kernel.

To reduce unwanted noise, the values are filtered by their standard deviation, and pixels with values less than 1.5 times the standard deviation are assigned a value of 0, keeping only the extreme pixels representing motion. After noise reduction, the next step is to find the boundaries of the motion. To do so, we create contours [232], which are short curves that connect points of the same hue or intensity. We select the two most significant contours in each frame to create a mask that indicates the motion area in a frame of a specific video. The main reason for choosing two contours is that in our datasets, an action is defined by the hands and the corresponding object. We create a bounding box around the resulting area that precisely represents the motion in each video. In fig. 3.9(a), we qualitatively compare our automatic box predictions and the provided supervised annotation for EPIC-KITCHENS-100 for several sample frames.

3.1.2 Motion-focused Self-Supervised Learning

MOFO uses 3D tube volume embeddings for the self-supervised pretext stage to obtain 3D video patches from frames as inputs. It encodes these with a vanilla ViT [59] with joint space-time attention as a backbone. We segmented each video into N non-overlapping tubes $\mathbf{p}_i \in \mathbb{R}^{H_t \times W_t \times T_t}$. Then, we use a high-ratio tube masking approach to perform masked autoencoder (MAE) pretraining with an asymmetric transformer-based encoder-decoder architecture reconstruction task. Unlike other often random masking methods, we explicitly integrate the motion information computed in subsection 3.1.1 into our masking strategy, resulting in a motion-guided approach to encode motion for our MAE. Our novel tube masking strategy enforces a mask to be applied to a high proportion of the tubes inside the motion area. Masking during self-supervised pretraining is applied across the entire spatiotemporal video, rather than being restricted to motion regions, with motion emphasis introduced by enforcing that a fixed proportion of the masked tubes lie within the automatically detected motion areas. In other words, a fixed percentage of the tubes (generally 75%) inside the motion area is always randomly masked to ensure the model attends more to the motion area at reconstruction time. Therefore, we apply an extremely high ratio masking at random (90%) while always masking a fixed percentage of the tubes (75%) inside the motion area. The encoder produces a latent feature representation of the video using input frames with blacked-out regions. The decoder uses the latent feature representation from the encoder and estimates the missing region using the mean squared error (MSE) loss, which is computed in pixel space between the masked patches and trained reconstructed outputs. Our design encourages the network to capture more useful spatiotemporal structures, making MOFO a more meaningful task and improving the performance of self-supervised pretraining. All models use only the unlabelled data in each dataset's training set for pretraining.

3.1.3 Motion-focused Finetuning

Recall that the self-supervised learning protocol is split between a pretraining and finetuning stage. We propose a new approach to focus on the motion area at both the pretext and the finetuning of the model. To do so, we utilise the trained asymmetric transformer-based encoder-decoder architecture for the video self-supervised pretraining and replace the encoder.

For each video, the region with motion is identified, and a bounding box for this area of interest is generated without supervision in subsection 3.1.1. As the area inside the motion box contains more semantic motion information, we wish to exploit this information for our task by leveraging the detected motion box. On the other hand, the video's setting and any nearby items could provide context for categorising the video clips for the action recognition task. For instance, in the case of washing dishes, the hands can be seen in the sink, but the dishes beside the sink may indicate that the person is washing them. To avoid missing motion area information while maintaining context from the outer area, we propose using multi-cross attention (MCA) [184] in our encoder. MCA is an attention mechanism that combines two distinct embedding sequences from the same modality. Unlike self-attention, where the inputs are the same set, during cross-attention, the inputs come from different sources. MCA extends this mechanism by computing attention scores across embeddings from multiple information sources. This module resides between the encoder and MLP classifier layers, takes the inner and outer motion box embeddings, and outputs the fused embedding. Here, the inner and outer regions are defined as spatio-temporal tubes rather than single-frame boxes: the inner tube follows the dominant motion across frames, while the outer tube is its complementary contextual region, which differs from R*CNN [85] by using motion-driven tube proposals and attention-based fusion instead of proposal-based region selection in static images. This module aims to integrate the video context while focusing on inner motion box embeddings by fusing outer and inner motion box embeddings. The outer region is defined as the complementary set of spatiotemporal patches that lie outside the automatically detected motion box and represent the surrounding contextual background information. Given a set of patches $\{\mathbf{p}_i\}_1^N$, the transformer yields two sets of embeddings: $\{\mathbf{e}^{\text{inner}}\}_{j=1}^{N_{\text{inner}}}$ for the inner motion boxes and $\{\mathbf{e}^{\text{outer}}\}_{k=1}^{N_{\text{outer}}}$ for the outer ones, as described by:

$$\{\mathbf{e}^{\text{inner}}\}_{j=1}^{N_{\text{inner}}}, \{\mathbf{e}^{\text{outer}}\}_{k=1}^{N_{\text{outer}}} = \text{ViT}\big(\{\mathbf{p}_i\}_1^N\big) \tag{3.3}$$

These embeddings are then processed by a cross-attention mechanism, where Q, K, and V represent query, key, and value, respectively. The CrossAttention function is formalised as follows:

$$\text{CrossAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.4}$$

where $Q = \mathbf{e}^{\text{inner}}$, $K = V = \mathbf{e}^{\text{outer}}$. In the context of multi-head attention, each attention head i is computed by applying the CrossAttention function to the query, key, and value matrices, each weighted by a different learnt weight matrix, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}, W_i^k \in \mathbb{R}^{d_{model} \times d_k}, W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ respectively:

$$\text{head}_i = \text{CrossAttention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3.5}$$

Finally, the fused embedding $\mathbf{e}^{\text{fused}}$ is computed by concatenating the results from all attention heads and then applying another learned weight matrix $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. The matrix W^O is a trainable output projection that is learned end-to-end during training to map the concatenated attention heads back to the model embedding space. This multi-head cross-attention (MCA) operation can be represented as:

$$\mathbf{e}^{\text{fused}} = \text{MCA}(Q, K, V) = \text{Concat}\left(\text{head}_1, \cdots, \text{head}_h\right) W^O \tag{3.6}$$

We employ $h = 3$ parallel attention layers, or heads, in this work. We also use $d_v = d_k = d_{model}$ for each. The model is ultimately finetuned with a cross-entropy loss $\mathcal{L}$:

$$\begin{aligned} \mathcal{L} &= -\sum_n \mathbf{y}_n \log \hat{\mathbf{y}}_n \\ \hat{\mathbf{y}} &= \text{FC}(\mathbf{e}^{\text{fused}}) \end{aligned} \tag{3.7}$$

where, $\mathbf{y}_n$ is the true label for nth video clip, $\hat{\mathbf{y}}_n$ is its predicted label, and FC is the fully connected layers typically used for classification.

3.2 Action Recognition Datasets

Human action recognition aims to understand human activities occurring in a video so that humans can understand. While some simple actions, like standing, can be recognised from a single frame (image), most human actions are much more complex and occur over a longer period of time; therefore, they must be observed through consecutive frames (video). To assist organisations in understanding real-time action and dynamic, organic movement, AI/ML models use datasets of human actions.

Something-Something V2 (SSV2) [87] The Something-Something V2 dataset is a large-scale, publicly available benchmark designed for fine-grained human–object interaction recognition. Unlike conventional action recognition datasets that focus primarily on scene or object semantics, SSV2 was specifically designed to capture subtle motion-based distinctions between similar actions. The dataset was collected through extensive crowdsourcing, with numerous participants recording short video clips of everyday physical interactions with objects. Each action prompt follows a template-based structure, such as "Putting [something] on [something]" or "Moving [something] from left to right", allowing workers to fill in the placeholders with arbitrary objects.

In total, the dataset contains 220,847 video clips spanning 174 action classes, with varying spatial resolutions and temporal lengths, typically lasting 2 to 6 seconds. The dataset is divided in an 8:1:1 ratio, consisting of 168,913 videos in the training set, 24,777 in the validation set, and 27,157 in the test set. The dataset emphasises third-person viewpoint recordings, in which only the human hands and the manipulated objects are visible This dataset captures subtle differences in motion and object manipulation rather than focusing on object categories or scene context. SSV2 is particularly challenging because many clips involve the same objects and visually similar motions that correspond to different action categories, requiring models to distinguish actions primarily based on motion and intent rather than appearance. These characteristics make SSV2 a critical benchmark for evaluating a model's capacity to capture temporal dependencies, motion dynamics, and contextual reasoning in third-person video understanding.

EPIC-KITCHENS-100 [47] Egocentric vision, also known as first-person vision, is a subfield of computer vision that deals with the analysis of images and videos captured by a wearable camera, often worn on the head or chest, which naturally approximates the wearer's visual field. The idea of using egocentric videos has just recently begun to be utilised thanks to novel, lightweight and affordable devices such as GoPro and similar [187]. As a fundamental problem in egocentric vision, one of the tasks of egocentric action recognition aims to recognise actions of camera wearers from egocentric videos. This community did not have a very large dataset for pertaining or a common benchmarking dataset until the appearance of EPIC-KITCHENS [44, 45, 46].

The EPIC-KITCHENS-100 (EK 100) dataset is one of the largest and most complete egocentric datasets. It comprises 700 variable-length videos, totalling over 100 hours and approximately 20 million frames, captured in natural kitchen environments. Each video is divided into short

action segments with an average duration of 3.12 seconds, annotated with verb–noun pairs that describe the performed action, such as open cupboard. The dataset contains 97 verbs, 300 nouns, and 3,806 action classes, with 67,217 training and 9,668 test segments. EK100 presents several challenges, including limited field of view, occlusions, irrelevant objects, background clutter, and camera motion, all of which make action understanding highly complex.

Fig. 3.2 presents sample frames from the Something-Something V2 and EPIC-KITCHENS-100 datasets, highlighting the contrast between third-person and egocentric viewpoints. SSV2 focuses on controlled object manipulation in a fixed-camera setup, whereas EPIC-KITCHENS captures natural, first-person interactions within real kitchen environments. Together, SSV2 and EK100 provide a balanced experimental foundation, covering both third-person and first-person perspectives of human–object interaction.

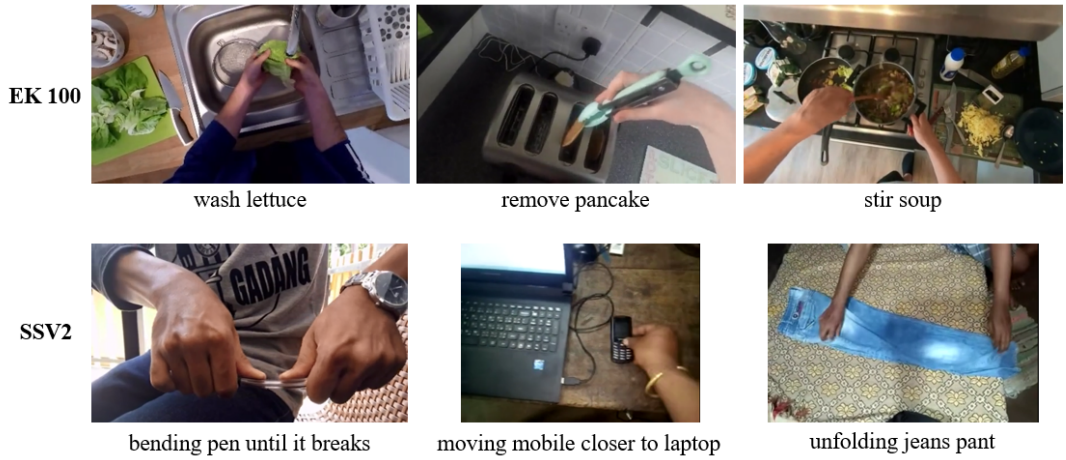


Figure 3.2: Sample frames and corresponding action classes from the datasets used in this work, illustrating the contrast between third-person and egocentric viewpoints in human–object interactions.

3.3 Experimental Setting

We use the technique of [260] to extract optical flow from a video to create a motion map, achieving 40% faster performance by parallelising I/O and computation.

MOFO employs ViT-Base as a decoder/encoder backbone, trained for 800 epochs on Something-Something V2 [87] and EPIC-KITCHENS-100 [47] for the SSL independently. We follow

the training and experiential parameters from recent work [238] to ensure a fair comparison and finetune for 100 epochs with early stopping. The model takes 16 frames from the video with a spatial resolution of 224×224 and divides the input video into 3D patch embeddings of $16 \times 16 \times 8$, resulting in $H = 224$, $W = 224$, $T = 16$, $H_t = 16$, $W_t = 16$, $T_t = 8$, and $N = 392$. Although the number of input patches is fixed, the number of inner (N_{inner}) and outer (N_{outer}) embeddings varies according to the size of the detected motion region in each video clip. The computational cost of MOFO is approximately 57 GFLOPs per video clip, which is in line with ViT-Base VideoMAE-style architectures under the same 16-frame, 224×224 input configuration, suggesting that the motion-focused design does not substantially increase computational complexity.

All experiments are conducted using PyTorch with DeepSpeed [133] on four NVIDIA Quadro RTX-5000 GPUs. We report Top-1 accuracy on EPIC-KITCHENS-100 and both Top-1 and Top-5 accuracy on Something-Something V2 for downstream action recognition tasks.

3.4 Results

The recognition accuracy for our MOFO SSL using regular finetuning is reported in table 3.1 shown as MOFO*. We demonstrate significant performance improvement over the other self-supervised approaches, comparable to the best supervised approach. Our strategy outperforms approaches like OmniMAE [82], trained jointly on images and videos, by 3.2% Top-1 accuracy. On Something-Something V2, our method outperforms VIMPAC [234] and ST-MAE [67], which both use ViT-Large as a backbone, whereas our backbone is vanilla ViT-Base with over 3x fewer parameters. Compared to VideoMAE [238], our approach achieves significantly better results while keeping the number of backbone parameters constant. While MOFO** indicates our result with pretraining on non-motion SSL and MOFO finetuning, which further increases accuracy, MOFO† denotes the MOFO SSL and MOFO finetuning, and this provides the greatest performance by increasing 2.6%, 2.1%, 1.3% accuracy over the best-performing methods on EPIC-KITCHENS-100 verb, noun, and action classification, respectively.

Table 3.1: Human activity recognition on **EPIC-KITCHENS** and **Something-Something** in terms of Top-1 and Top-5 accuracy. blue: This is the result computed by us using the public code MOFO* is pretrained by our MOFO SSL and uses non-motion finetuning. MOFO** This is our result with pretraining on non-motion SSL and has MOFO finetuning. MOFO† denotes the MOFO SSL and MOFO finetuning.

Method	Backbone	Param	**Something-Something** Action		**EPIC-KITCHENS**		
					Verb	Noun	Action
			Top-1	Top-5	Top-1	Top-1	Top-1
Supervised							
TDN_{EN} [253]	ResNet101×2	88	69.6	92.2	-	-	-
SlowFast [68]	ResNet101	53	63.1	87.6	65.6	50.0	38.5
TSM [149]	ResNet-50	-	63.4	88.5	67.9	49.0	38.3
MViTv1 [64]	MViTv1-B	37	67.7	90.9		-	-
TimeSformer [17]	ViT-B	121	59.9	-	-	-	-
TimeSformer [17]	ViT-L	430	62.4	-	-	-	-
ViViT FE [6]	ViT-L	-	65.9	89.9	66.4	56.8	44.0
Mformer [192]	ViT-B	109	66.5	90.1	66.7	56.5	43.1
Mformer [192]	ViT-L	382	68.1	91.2	67.1	57.6	44.1
Video SWin [166]	Swin-B	88	69.6	92.7	67.8	57.0	46.1
Self-supervised							
VIMPAC [234]	ViT-L	307	68.1	-	-	-	-
BEVT [259]	Swin-B	88	70.6	-	-	-	-
VideoMAE [238]	ViT-B	87	70.8	92.4	71.6	66.0	53.2
ST-MAE [67]	ViT-L	304	72.1	-	-	-	-
OmniMAE [82]	ViT-B	87	69.5	-	-	-	39.3
Omnivore(Swin-B) [84]	ViT-B	-	71.4	93.5	69.5	61.7	49.9
Ours(MOFO*)	ViT-B	87	**72.7**	**94.2**	**73.0**	**67.1**	**54.1**
Ours(MOFO)**	ViT-B	102	**74.7**	**95.0**	**74.0**	**68.0**	**54.5**
Ours(MOFO†)	ViT-B	102	**75.50**	**95.3**	**74.2**	**68.1**	**54.5**

3.4.1 Visualising self-supervised representation

To further understand the representations learnt by MOFO, we utilise GradCAM [212] to create a saliency map highlighting each pixel's importance to show how each pixel contributes to the discrimination of the video clip. Fig. 3.3 visualises the middle frame of a video clip, the corresponding motion map and the Grad-CAM attention maps for VideoMAE and our MOFO. It is interesting to note that for similar actions: *knead dough*, *cut carrot*, and *cut-in tomato*, MOFO is sensitive to the most significant motion location as detected by our automatic algorithm. Additionally, in fig. 3.3's middle row, the MOFO model not only detects the motion area but also highlights the direction of the hand movement. The representations learnt by MOFO align the motion map more effectively than those learnt by VideoMAE. Additionally, the regions of high

importance in these representations serve as solid visual evidence for describing the video clip. As a result, the representations learnt by MOFO may be more robust. These findings further demonstrate the effectiveness of utilising motion in MOFO for SSL.

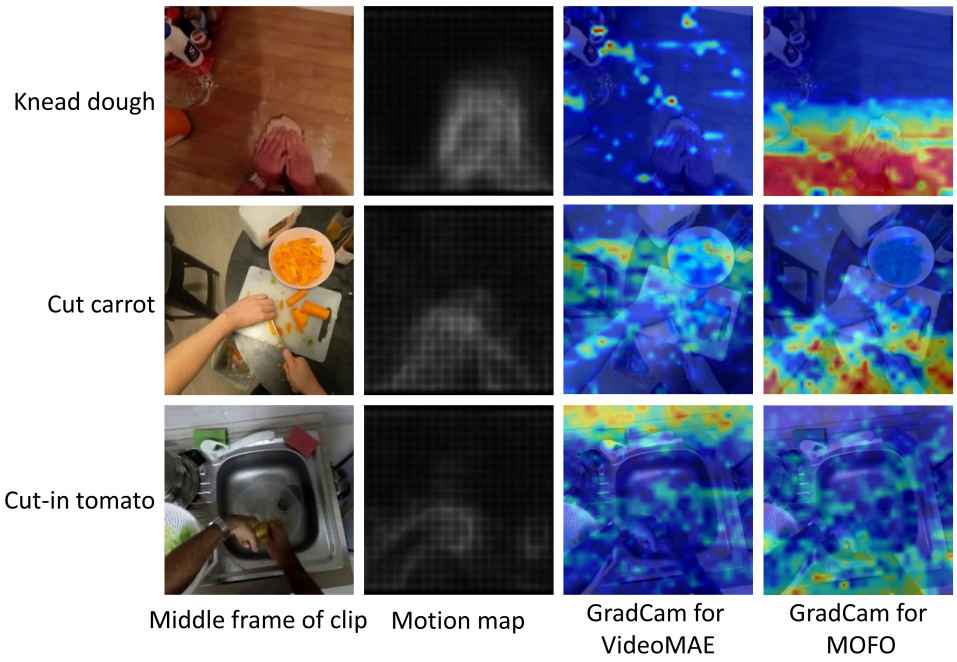


Figure 3.3: From left to right: video clip middle frame, motion map, Grad-CAM attention map for VideoMAE, Grad-CAM attention map for MOFO

3.4.2 MOFO Reconstruction Results

This section shows several reconstructed image frames from a video in fig. 3.4 and fig. 3.5. With tube masking, the model learns to reconstruct masked patches by leveraging spatial and temporal information from the adjacent unmasked regions across frames. The mean squared error (MSE) loss between normalised masked tokens and reconstructed tokens in pixel space serves as the loss function. Videos are randomly selected from the validation sets of both datasets. Unlike the VideoMAE model, our proposed MOFO model enforces a fixed proportion of masking within the detected motion areas, ensuring consistent emphasis on motion regions. These examples suggest that, compared to VideoMAE, our MOFO model reconstructs samples in the motion area significantly more accurately, demonstrating that it has focused its attention on this area. We can achieve satisfying reconstruction results, especially in the area where motion occurs,

with our MOFO by applying an extremely high ratio masking at random (90%) while always masking a fixed percentage of the tubes (75%) within the motion area.

3.4.3 Visualization of GradCAM using MOFO self-supervision

Fig. 3.6 visualises the GradCAM and motion map for the samples in which VideoMAE can't identify the class, but our MOFO can. The attention maps show how effective our approach is in capturing the motion area.

3.5 Automatic Motion Area Detection

In fig. 3.7, we present additional qualitative examples of our automatic motion area detection compared with the provided supervised annotation for the EPIC-KITCHENS-100 and Something-Something V2 datasets. These samples show that our proposed automatic motion area detection minimises the impact of the static object in the motion box while highlighting the motion areas. Our automatic motion box concentrates on the area and item of interest, which is a crucial necessity for our proposed approach, even for self-supervision or finetuning.

3.5.1 Ablation studies

We evaluate the contribution of the different components of our approach.

Masking ratio

VideoMAE [238] recommended tube masking with an extremely high ratio, which helps reduce information leakage during masked modelling. They demonstrated the best efficiency and efficacy with a masking ratio of 90%. Therefore, we explore the effect of the inside masking ratio for verb classification on EPIC-KITCHENS-100 in fig. 3.8. It shows that the model pretrained with a masking ratio of 90% as the general masking ratio for a video and a high ratio for the inside masking ratio (75%) achieves the highest efficiency level. Thus, we continue experimenting with the rest by fixing the inside mask ratio to 75%.

Automatic vs. supervised motion area detection

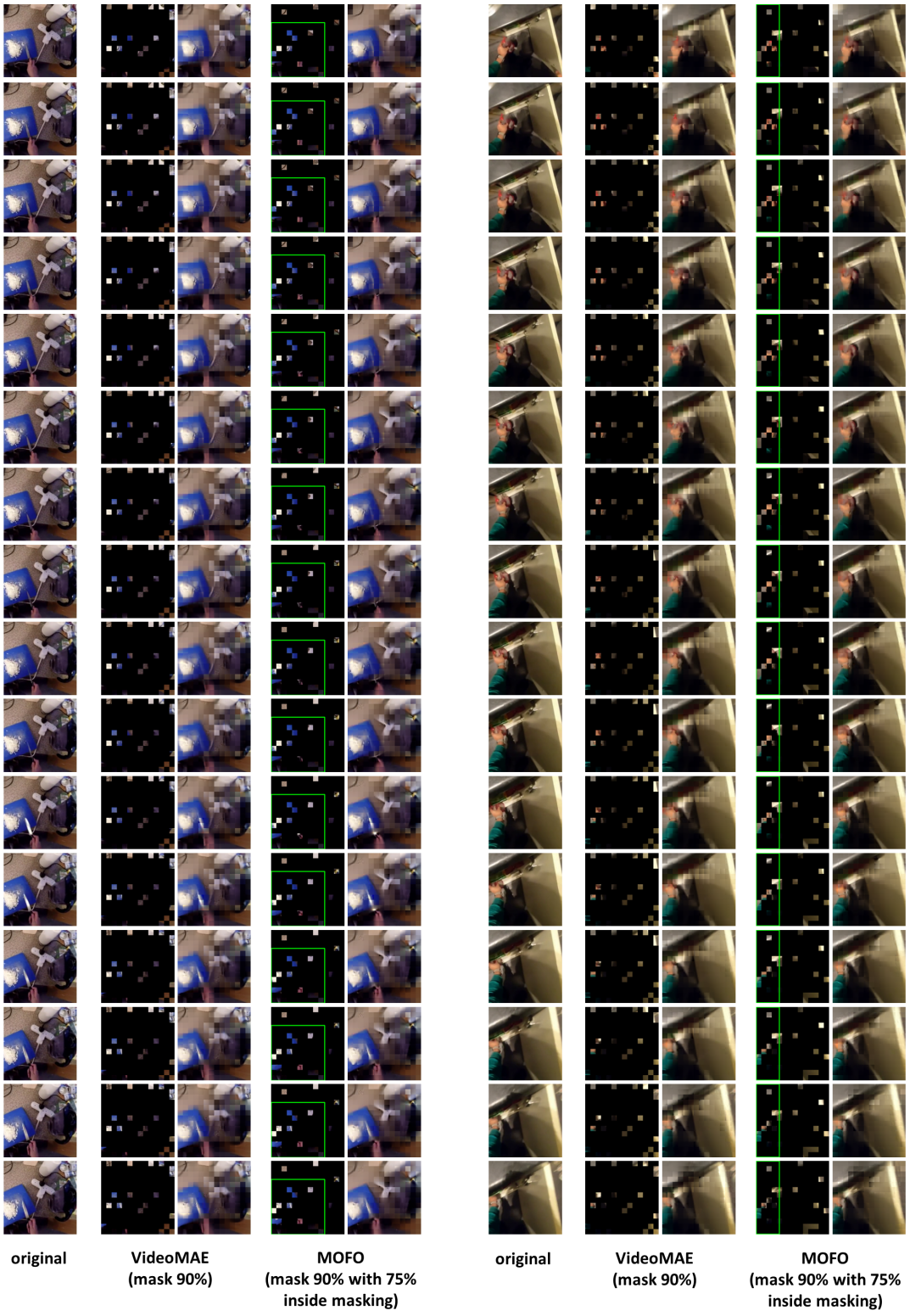


Figure 3.4: Qualitative comparison of reconstructions using VideoMAE and MOFO on **EPIC-KITCHENS-100** dataset. MOFO Reconstructions of videos are predicted by MOFO pretrained with a masking ratio of 90% and an inside masking ratio of 75% .

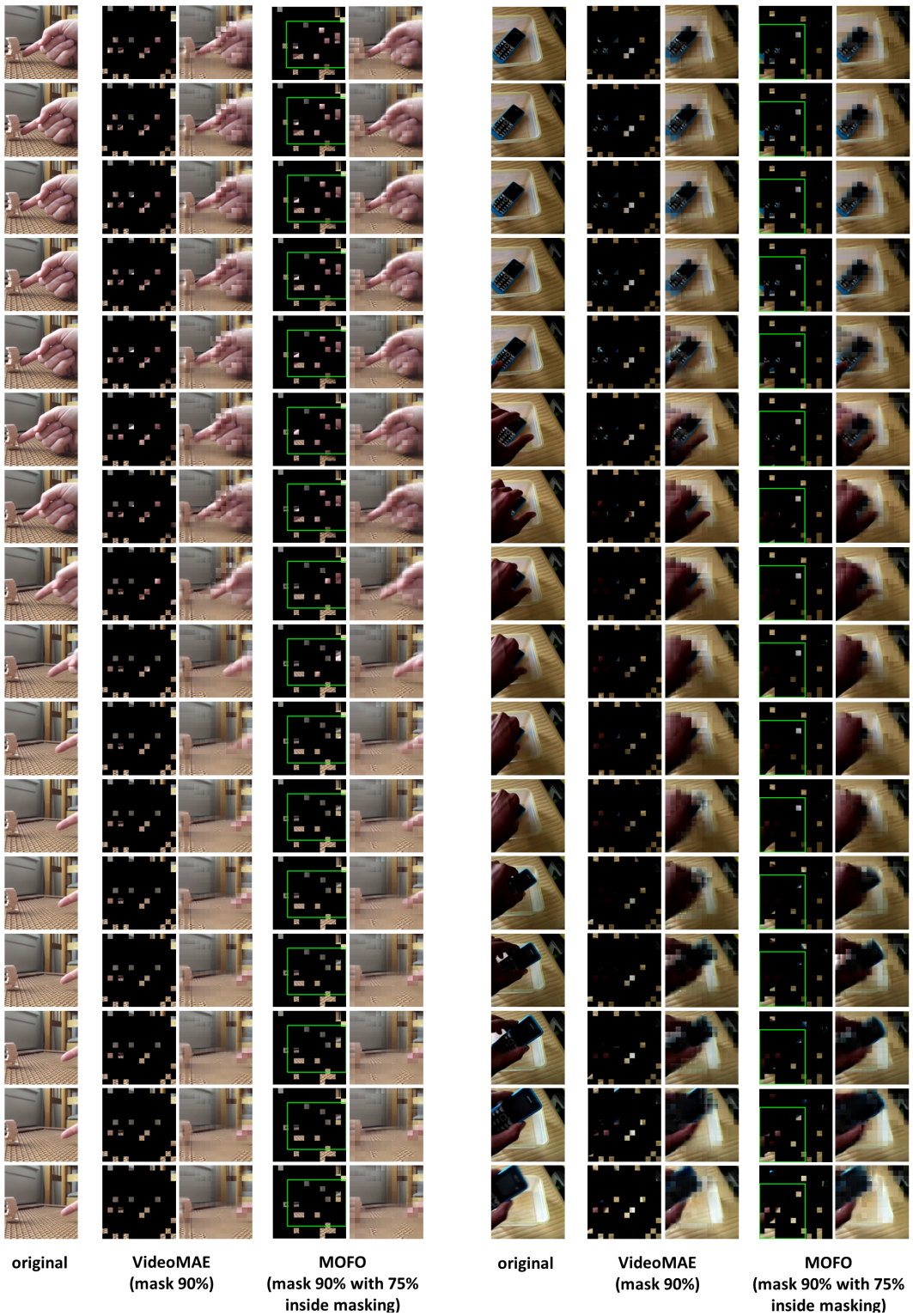


Figure 3.5: Qualitative comparison of reconstructions using VideoMAE and MOFO on **the Something-Something V2** dataset. MOFO reconstructions of videos are predicted by MOFO pretrained with a masking ratio of 90% and an inside masking ratio of 75%.

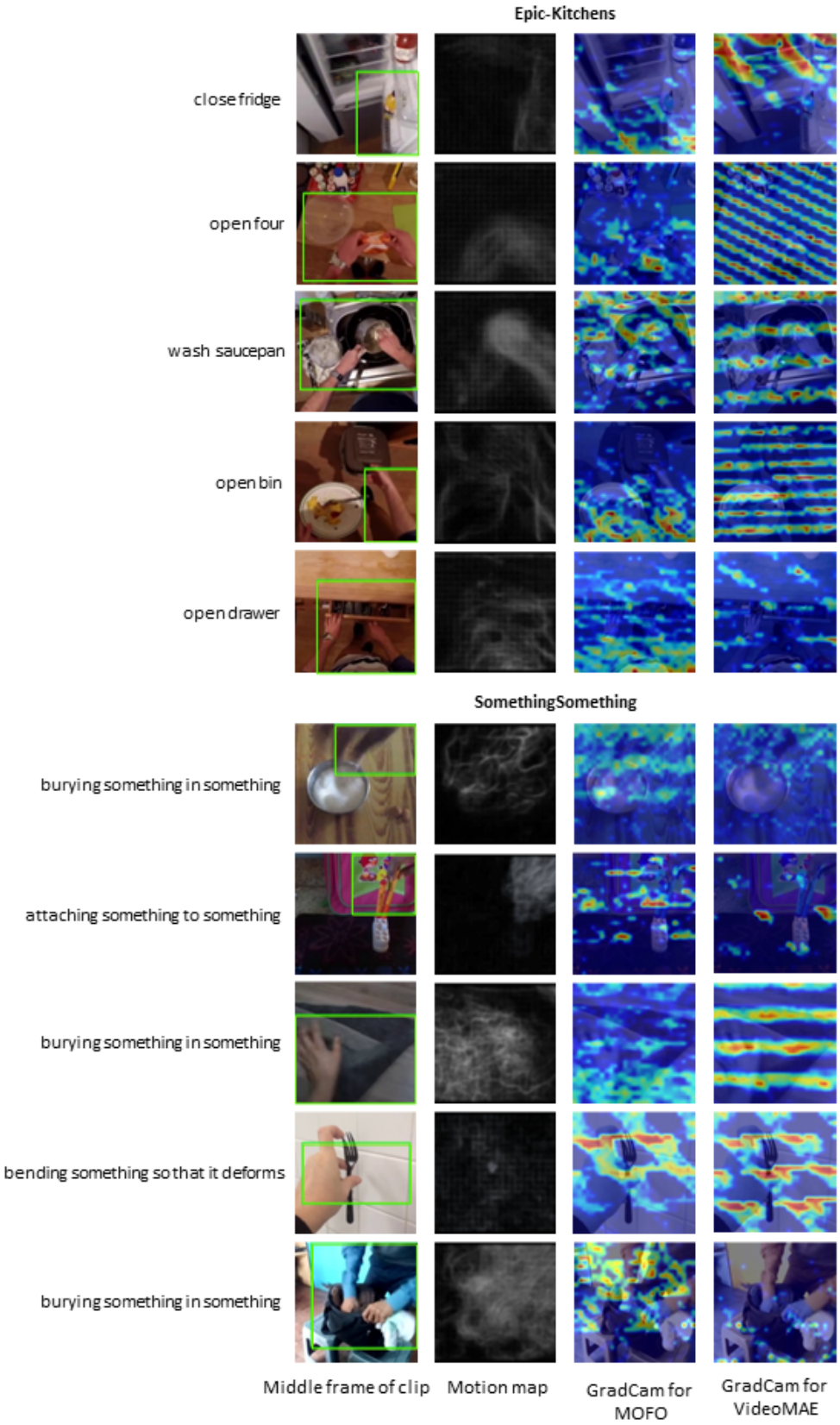


Figure 3.6: We visualise the attention maps generated by GradCAM based on VideoMAE and MOFO for the EPIC-KITCHENS-100 and Something-Something V2 datasets. The attention maps show that our proposed approach can better capture the motion area.

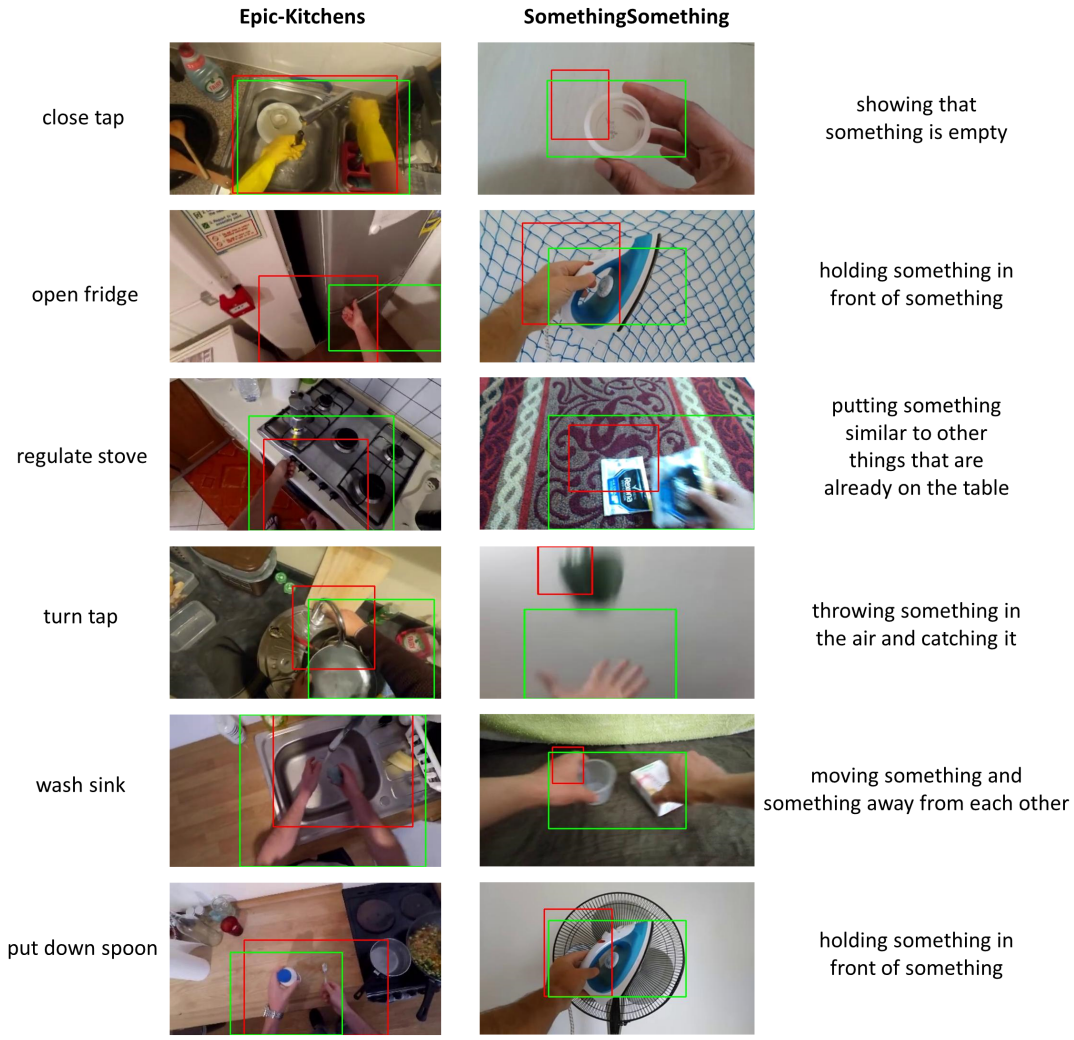


Figure 3.7: Comparison between the unsupervised and supervised motion area detection, green rectangles indicate the unsupervised, while red ones show the supervised detected motion area.

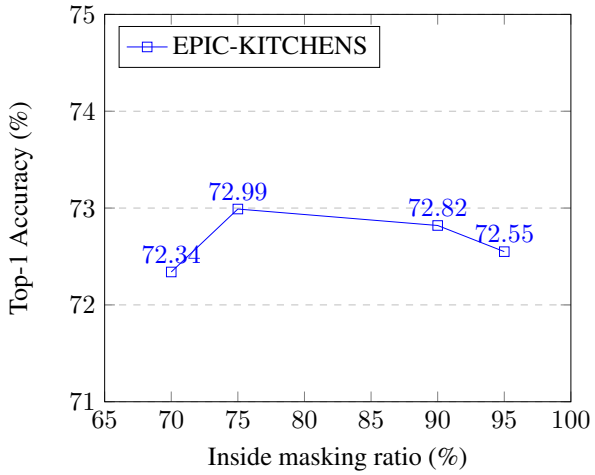


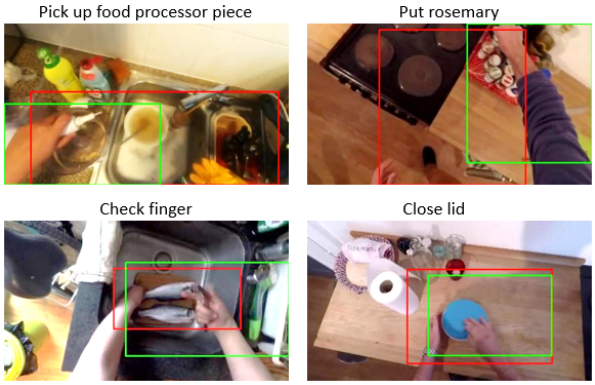
Figure 3.8: The effect of the inside masking ratio on EPIC-KITCHENS-100 dataset for verb classification demonstrates that a high inside masking ratio (75%) delivers the best efficiency and effectiveness trade-off.

We compare the results using our automatically detected motion areas and the ground truth bounding box annotation provided by [47] on the EPIC-KITCHENS-100 dataset in table 3.9(b). Our automatic motion detection results are close compared to supervised annotations, as seen in table 3.9(b), despite the challenging camera motion from the egocentric videos.

We compute the Intersection over Union (IoU) metric to compare our automatic detector with the supervised annotated bounding boxes on both datasets. For the EPIC-KITCHENS-100 dataset, the IoU is 40%, and for Something-Something V2, the IoU is 31%. Although these numbers are lower, our automatic motion detection only detects motion and ignores unnecessary static objects near the motion. As you can see in fig. 3.9(a), our automatic motion box still focuses on the motion area and object of interest, which is the key requirement.

MCA Hyper-parameters

We list the MCA hyperparameters used in our MOFO finetuning experiments here. We experiment with various head and depth settings when EPIC-KITCHENS-100 is the target dataset shown in table 3.2. We experiment with these parameters for the verb task on EPIC-KITCHENS-100 to find the best choice for our cross-attention layer that we suggested for MOFO finetuning. The final head and depth are 3 and 1, respectively.



(a)

Method	Annotation	EPIC-KITCHENS-100 Verb Top-1
MOFO supervision	Supervised	73.26
	Automatic(ours)	72.99

(b)

Figure 3.9: (a) Comparison between the unsupervised and supervised motion area detection, green rectangles indicate the unsupervised, while red ones show the supervised detected motion area. (b) Effect of supervised vs. automatic motion area utilisation in MOFO.

Table 3.2: Ablation experiment for the number of heads and depth in MOFO finetuning

Finetuning method	Backbone training	CA heads	CA depths	EPIC-KITCHENS-100 Verb Top-1
VideoMAE	VideoMAE	-	-	71.6
MOFO	VideoMAE	1	1	73.5
MOFO	VideoMAE	1	2	73.8
MOFO	VideoMAE	1	3	73.6
MOFO	VideoMAE	2	1	73.7
MOFO	VideoMAE	2	2	73.3
MOFO	VideoMAE	**3**	**1**	**74.0**
MOFO	VideoMAE	3	2	73.5
MOFO	VideoMAE	4	1	73.8
MOFO	VideoMAE	4	2	73.3

3.6 Conclusion

This chapter introduced MOFO, a motion-focused technique that explicitly incorporates motion information to enhance motion-aware self-supervised video action recognition. Unlike traditional SSL methods that often overlook motion or rely on computationally expensive annotations, MOFO leverages motion boundaries derived from motion maps to identify salient dynamic regions in videos. This motion-aware approach is seamlessly integrated into both the pretraining and finetuning stages, enabling the model to focus on motion-rich areas while maintaining contextual awareness of the surrounding environment.

The core contributions of MOFO lie in its innovative masking strategy and multi-cross attention mechanism. During pretraining, MOFO employs a motion-guided masking approach within a masked autoencoder architecture, prioritising the reconstruction of motion-dense regions. This ensures that the model learns spatiotemporal representations that are highly relevant to action understanding. In the finetuning stage, MOFO further emphasises motion by fusing embeddings from both within and outside the detected motion areas via multi-head cross-attention. This dual-stage focus on motion enables MOFO to capture both fine-grained dynamics and broader contextual cues, resulting in more robust video representations.

Our extensive experiments on two challenging datasets, Something-Something V2 (SSV2) and EPIC-KITCHENS-100, demonstrate the effectiveness of MOFO. The proposed framework significantly outperforms state-of-the-art SSL methods, achieving a 4.7% improvement in Top-1 accuracy on SSV2 and notable gains across verb, noun, and action classifications on EPIC-KITCHENS-100. These results highlight the importance of explicitly encoding motion in SSL frameworks for video understanding tasks. Additionally, our method achieves these improvements with fewer parameters than some competing approaches, demonstrating its efficiency.

Beyond its empirical performance, MOFO also addresses key challenges in video action recognition, such as mitigating the effects of camera-induced noise and reducing dependency on costly annotations. By focusing on motion boundaries rather than raw optical flow or object detection outputs, MOFO provides a robust and scalable solution for real-world applications where labelled data is scarce or camera movement is prevalent.

In conclusion, MOFO represents a significant step forward in self-supervised video understanding by introducing a motion-focused paradigm that bridges the gap between spatiotemporal reasoning and action recognition. Its ability to learn from unlabelled data while emphasising dynamic regions opens up new possibilities for SSL research and practical applications in domains such as surveillance, robotics, and human-computer interaction. Future work could extend this framework to other video understanding tasks or integrate it with multimodal data to yield even richer representations.

Currently, MOFO focuses exclusively on visual and motion cues extracted from video data, which limits its ability to capture semantic context or high-level descriptions of actions. Incorpo-

rating text-based annotations, such as subtitles, captions, or action descriptions, could provide complementary information that enhances the model's understanding of complex actions and their associated contexts. So, one potential area for improvement in MOFO is the integration of textual information as an additional modality to enrich the self-supervised learning process and address limitations in semantic reasoning, enhancing MOFO's ability to recognise complex actions in diverse contexts. In the next chapter, we introduce a method for integrating language supervision directly into the feature learning process to create more semantically robust video models.

Chapter 4

FILS: Self-supervised Video Feature Prediction In Semantic Language Space

In the previous chapter, we focused on capturing motion-aware visual representations by explicitly emphasising motion cues through automatic motion area detection, self-supervised motion-guided masking, and subsequent finetuning. While this approach effectively captured temporal and motion cues, it remained limited in its ability to encode higher-level semantic information and contextual understanding of actions. To address this limitation, this chapter explores integrating language-based supervision into self-supervised video learning to enrich video representations with semantic meaning. We propose FILS, a framework that predicts video features within a semantic language space, bridging visual representation learning with language-grounded semantics to enable richer and more meaningful representation learning.

The rapid evolution of self-supervised learning has revolutionised the field of video representation, enabling models to learn meaningful features from unlabelled data without relying on costly human annotations. Mask-based reconstruction techniques, initially successful in natural language processing [52, 20] and later extended to the visual domain [238], have shown great promise for learning robust representations. In parallel, leveraging web text as an addi-

tional source of self-supervision for visual learning has shown promise [200], with applications extending to video [322].

Building on this motivation, video understanding presents unique challenges compared to static image analysis, as videos are characterised by dense temporal data, frame-to-frame redundancy, and critical action segments essential for comprehension. Traditional approaches borrowed from image-based techniques often fail to address these nuances, resulting in suboptimal performance in video tasks. Building on the motion-aware masking strategy introduced in the previous chapter, we extend this idea by integrating language-based supervision to achieve semantically grounded video representation learning.

Related generative approaches explore selective vision–language alignment through reconstruction or synthesis objectives, such as text-conditioned video generation [174]; however, FILS differs by focusing on motion-aware selective contrastive learning for representation learning rather than generative reconstruction.

Addressing text supervision, early approaches utilised labelled text from supervised datasets, limiting models to predefined categories and labels [257]. Recent advancements have shifted towards leveraging Large Language Models (LLMs) and vision-language models for text supervision, either as a text bag [155] or densely captioned videos [322]. Leveraging textual information for supervision has gained traction in multimodal learning, where models align visual and linguistic features to enhance semantic understanding. To effectively apply text supervision, contrastive learning on related images and captions has been shown to build powerful representations, demonstrating that language-supervised visual pretraining, such as CLIP [200], is a simple yet effective technique for learning representations. Rapid progress has been achieved due to the effectiveness and ease of use of contrastive learning approaches on images [32, 182, 57]. These approaches lay the foundation for effectively combining text supervision with self-supervised learning; however, directly applying these image-based techniques to videos introduces complexities due to the temporal dimension and the need to focus on action-relevant frames. In an attempt to transfer image-based CLIP to the video domain, several works use specific (or all) frames and compute the average of their features to provide a representation of the video clip [185, 155] as videos are computationally intensive and intrinsically complex. Nevertheless,

their counterparts in the video domain do not show the same generality [197, 257] and also do not take temporality into account, a key aspect for understanding a video.

Defining the objective for reconstructing the masked video presents another challenge. The original masked autoencoder work did reconstruction in pixel space [238]. Tan et al. [234] note that such a reconstruction would make the model overfit on detailed low-level visual information and instead propose to reconstruct in the latent space of a quantised-variational autoencoder. Similarly, Assran et al. [8], working on images, find that learning in a higher semantic space leads a model to learn more semantic features. Yang et al. [296] take this to language semantics, aiming to reconstruct image patches by mapping them to a distribution over textual features. They aim to predict the semantics of masked patches within the linguistic context. This approach involves two primary training objectives: first, employing image-text contrastive learning to unify the two modalities within a shared embedding space, and second, incorporating a reconstructive loss. However, their image-text contrastive learning involves randomly selecting image patches. While effective for certain tasks, this strategy may not be optimal for video representation learning, where only some part of the depicted scene captures the activity that should align with the video's text caption. Suppose we could focus contrastive learning on the parts of the video critical to understanding an activity. In that case, the video-text alignment can occur more efficiently, emphasising the essential semantic features necessary for comprehensive video understanding.

The impressive progress in these research directions encouraged us to consider whether masking and language guidance can be jointly leveraged to build richer, semantically grounded video representations. A straightforward approach to achieving this objective is to integrate a masking strategy with video-text contrastive learning for multi-task learning, but this simple combination cannot fully leverage the potential synergies between these two objectives. Self-supervised objectives operating within a latent space established by language can maintain the alignment of language with learnt visual representations. This enhances the interpretability and semantic aspects of the representations.

We address these combined issues by drawing on the best practices from related works and developing FILS to inherit the benefits of these approaches while addressing their limitations through an innovative strategy. The core idea is to predict the features of masked patches within

a semantic language space in a self-supervised manner. In other words, the model predicts semantic feature embeddings of masked patches in the shared vision–language space, rather than reconstructing their raw pixel values. Text representations act as prototypes for transforming vision features into a language space, enabling semantically meaningful feature prediction. This approach not only enhances interpretability but also improves the semantic richness of learnt representations.

For text supervision, we use the natural descriptions generated by an off-the-shelf video captioning model [302]. Unlike image-based vision–language models such as BLIP and BLIP2 [137, 136], which are primarily designed for static image–text alignment, or grounding-focused models such as Grounding DINO and DINOv2 [22, 189], VideoBLIP [302] explicitly models temporal dynamics, making it more suitable for generating clip-level captions that describe actions unfolding over time.

We narrow our focus on the video regions where significant action unfolds. We detect this *action area* using a method that detects motion from optical flow derivatives while ignoring camera motion introduced as described in Chapter 3. Furthermore, FILS introduces ActCLIP, an auxiliary objective that performs contrastive learning between patches in recognised action areas and their corresponding generated text captions. In contrast to approaches that explicitly model verbs in the language stream, such as Visual-Dynamic Injection (VDI) [175], ActCLIP does not isolate verbs during supervision. Instead, we use the full caption as weak semantic guidance, while action relevance is injected through the visual modality by selecting motion-salient patches. This integration of motion-based saliency with language guidance enhances the model's ability to capture the essence of depicted activities while maintaining computational efficiency. To learn high-level semantic features, we train our video encoder to learn embeddings in the CLIP-based language space. These are visual features projected into and structured within a language-aligned semantic space, enabling the model to exploit the semantic structure provided by language while remaining fundamentally grounded in visual information. ActCLIP ensures that the model focuses on motion-salient regions and aligns them with the language semantics of the video, capturing the key dynamics that convey the meaning of each activity. While contrastive learning benefits from larger batch sizes, our ActCLIP shows promise in surpassing CLIP's performance in video representation learning while maintaining low batch sizes and computation requirements. Subsequently, we then define our objective for predicting

masked video features within this semantic space. The same video encoder passes embeddings to a predictor that predicts masked embeddings in the language space.

In Figure 4.1, you can see how the proposed FILS architecture differs from existing methods. In (a), MAE performs video reconstruction through a masking and decoding process directly in the pixel domain. In (b), CLIP aligns video and text features using a contrastive learning objective. In (c), MAE+CLIP simply combines these two ideas by jointly applying masking-based pixel-domain reconstruction and video–text contrastive learning in parallel. Finally, (d) illustrates our FILS framework, which departs from pixel-level reconstruction and instead predicts features of masked video patches in the semantic language space. This process is guided by our proposed motion-aware contrastive learning (ActCLIP), which focuses on detected action areas to strengthen video–language alignment. The red arrows indicate the knowledge flow toward the language space, while the black arrows show the flow within the visual domain. We note that the teacher branch in Fig. 1(d) does not perform reconstruction; it is used only to provide target features for supervising the student through the feature prediction objective. The reconstruction step is carried out exclusively in the student path via the predictor, which estimates the masked patch features in the semantic language space. In parallel, the contrastive learning objective (ActCLIP) supervises the student video encoder by aligning video features with text embeddings.

Our method outperforms all previous self-supervised learning techniques using ViT-b and demonstrates excellent generalisation capability on a downstream task such as action recognition. Experiments demonstrate that it achieves state-of-the-art results while maintaining an efficient network design. With a vanilla ViT-B as the vision model, FILS achieves 51.0% and 72.1% top-1 accuracy on EPIC-KITCHENS-100 and Something-Something V2 action recognition, which are $+4.1\%$ higher than LAVILA [322]-Base version and $+0.9\%$ better than VideoMAE V2-Base [252], after finetuning, respectively. We observed a $+1.3\%$ improvement in mAP on Charades-Ego and a $+1.03\%$ increase in top-1 accuracy on EGTEA after finetuning, compared to the baseline presented in the LAVILA study [322]. The following briefly describes our primary contributions:

- A new self-supervised method for representation learning from unlabelled videos and their captions. A pre-task feature prediction strategy on masked video and patch-wise

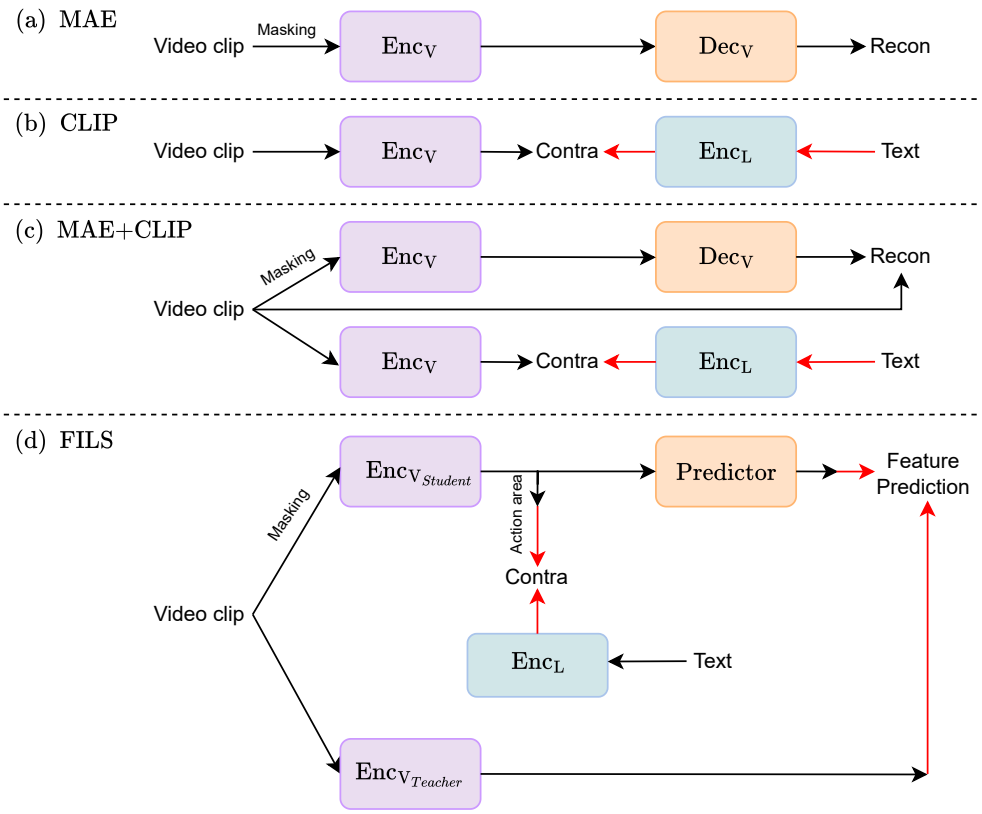


Figure 4.1: Architecture comparisons between MAE, CLIP, MAE+CLIP and FILS. Contra indicates video-text contrastive loss. The red arrow points to the language space, while the black ones indicate the knowledge flow in the vision space.

contrastive learning within potential action areas enriches our video encoder with richer, more abstract representations.

- Recognising the crucial importance of motion in self-supervised learning methods for identifying actions in videos, in our proposed ActCLIP, we actively integrate it by conducting contrastive learning between patches within the recognised action region and an associated textual context.

- Illustrating the efficacy of utilising mutual information to prioritise patches conveying semantic and action-relevant details, thereby enhancing the feature prediction process within the semantic language space for action recognition on challenging datasets like EPIC-KITCHENS-100, Something-SomethingV2, Charades-Ego, and EGTEA.

In summary, FILS represents a significant advancement in self-supervised video representation learning by merging masking strategies with language guidance in a unified framework. Its

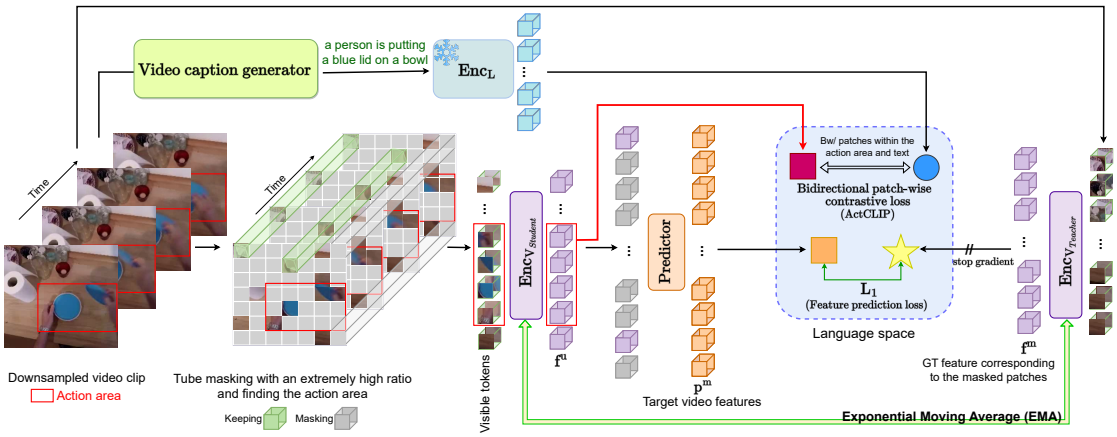


Figure 4.2: Overview of our method. We perform self-supervised feature prediction and video-text contrastive learning simultaneously. The red arrow denotes the features of the patches within the action area.

ability to focus on action-relevant regions while leveraging textual information positions it as a powerful tool for video understanding tasks. The following sections will delve into the technical details of FILS, its architecture, training objectives, and experimental results that underscore its effectiveness in achieving high performance in semantic video representation.

4.1 Methodology

We introduce FILS, a framework for deeper video understanding that leverages a unified embedding space that integrates video and natural language for multimodal learning. Fig. 4.2 presents an overview of our method; it has two objectives: 1) Feature Prediction, where the input is masked and encoded (the student mode with random initialisation), and the predictor predicts representations. Then, we create representations of all the input data, which are meant to be targeted for the learning task (teacher mode). As we will discuss in Sec. 4.1.1, to prevent the collapse, the teacher tracks student parameters, and their weights are derived from the exponentially moving average of the student [98, 89]. 2) ActCLIP, an auxiliary CLIP-based self-supervised objective performing contrastive learning between motion or action area patches and relevant text, aligning video and language spaces to learn semantic context. More information is in Sec. 4.1.2.

4.1.1 Model Architecture

Video Encoder. In our approach, we cast video representation learning as a feature prediction objective that compares predicted features from the predictor with their corresponding ground-truth features. To ensure stable representation during this task, we adopt a teacher/student approach. As is typical with self-supervised works [238, 252], we utilise an encoder ($\text{Enc}_{\text{V}_{Student}}$) to compute the representation of the masked input using the unmasked (remaining) patches. As tube modelling has demonstrated superior capability for capturing temporal and spatial information compared to frame modelling [238], we utilise tube embedding for a video clip to further concentrate on spatiotemporal saliency. The input to the student encoder is masked with a high proportion of video (V) patches using spatiotemporal tube masking. After masking, the student vision encoder encodes the unmasked patches (V_u), resulting in an f^u embedding. The ground truth of masked patches in the video input (V_m) is encoded by a teacher encoder, resulting in an f^m embedding. Eq. 4.1 formulates the process of vision encoders:

$$\begin{aligned} f^u &= \text{Enc}_{\text{V}_{Student}}(V_u), \qquad u \in [1, N_u] \\ f^m &= \text{Enc}_{\text{V}_{Teacher}}(V_m), \qquad m \in [1, N_m] \end{aligned} \tag{4.1}$$

m and u demonstrate the patch index related to masked and unmasked patches. N_u and N_m indicate the number of unmasked (visible) patches and the number of masked patches in that order. $N_u + N_m = N$ is the number of patches in the video.

Predictor. This module is made up of transformer blocks. Using a learnable [MASK] token and the unmasked encoded feature f_u of the masked video as inputs, it decodes or predicts the features of the masked patches from a masked view. This results in p^m, the predicted features of the masked input, where N_u indicates the number of unmasked (visible) patches.

$$p^m = \text{Predictor}(f^u, [\text{MASK}]), \qquad u \in [1, N_u], \tag{4.2}$$

Teacher Parameterisation. To prevent representation collapse, the teacher model is parameterised using an exponentially moving average (EMA) of the student model parameters θ, where the target weights Δ encode the video clip patches: $\Delta \leftarrow \tau\Delta + (1-\tau)\theta$. We employ a schedule for τ that increases this parameter linearly throughout the first τ_n updates, from τ_0 to the target

value τ_e. The value is then maintained constant for the rest of the training. When the model is random at the beginning of training, this strategy leads to more frequent updates to the teacher; however, when suitable parameters have already been learnt, the frequency of updates to the teacher decreases.

Text Encoder. To enable the integration of textual information with our vision encoder, we utilise a text encoder that converts textual inputs into a latent representation for joint processing. A stack of transformer layers tokenises the representation of input text (T) into the global representation of the input texts, $h = \text{Enc}_\text{L}(T)$.

4.1.2 Training Objectives

We propose two loss functions in our frameworks: a contrastive video-to-text loss and a feature prediction loss. These losses are crucial in aligning video and text modalities within the language embedding space, fostering effective cross-modal understanding.

ActCLIP. Recognising the importance of motion in self-supervised video techniques for action recognition, we integrate it through contrastive learning between patches within the motion or action area and relevant text, resulting in a unified language-vision space that fosters coherent content understanding. Motion areas are identified following an approach in Chapter 3 to represent motion precisely and are termed 'action areas' in our work. Then, to learn the parameters of the shared vision-language embedding space, we leverage a patch-wise video-text contrastive loss between patches within the action area to align both modalities in a shared embedding space. Given this, the patch-wise clip-level contrastive objective happens in the detected action area; we call this *ActCLIP*. Initially, we take the mean-pooled video feature of patches within the detected motion area:

$$\bar{f} = \frac{1}{N} \sum_{i=1}^{N_a} f^a \tag{4.3}$$

a indicates patches inside the action area, N_a is the number of patches in the detected action area, and f^a is the encoded feature representation of the patches in a. The video features are further mapped by the projection head to the language space, normalised, and the text feature

comes after:

$$z^V = \|\theta(\bar{f})\|, \qquad z^T = \|h\| \tag{4.4}$$

$\|.\|$ and $\theta(.)$ denote the normalisation operation and the projection head (mapping) from video to text, respectively.

The bidirectional contrastive loss between video and text can be represented as follows:

$$L_{V2T} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{exp(\langle z_i^V, z_i^T\rangle/\sigma)}{\sum_{j=1}^{B} exp(\langle z_j^V, z_j^T\rangle/\sigma)}, \qquad L_{T2V} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{exp(\langle z_i^T, z_i^V\rangle/\sigma)}{\sum_{j=1}^{B} exp(\langle z_j^T, z_j^V\rangle/\sigma)} \tag{4.5}$$

σ is a learnable parameter that is cooperatively trained throughout the pretraining; B denotes the batch size, and i and j represent the index inside a mini-batch. The following is a formulation of the total loss of video-text contrastive learning:

$$L_{ActCLIP} = \frac{1}{2}(L_{V2T} + L_{T2V}) \tag{4.6}$$

Feature Prediction in Language Space. Our reconstruction space is built by taking a vision-language perspective. We perform masked visual reconstruction in this language space, using the text features as natural semantic information for the video patches. Therefore, given the encoded patch features from the teacher encoder corresponding to the masked patches (f^m) and predicted patch features (p^m) (reconstructed ones), where m is the index of the masked patch, we first map and normalise both features to the language feature space:

$$\tilde{p}_i^m = \|\theta(p_i^m)\|, \qquad \tilde{g}_i^m = \|\theta(f_i^m)\| \tag{4.7}$$

where i indicates data in the batch and $\theta(.)$ represents the same vision mapping in Eq. 4.4. So, both the masked prediction and its corresponding target are mapped into this language semantic space. Our proposed loss function for the prediction is **F**eature **P**rediction (FP) loss, which is the average L1 distance between the mapped predicted patch-level representations via the predictor ($\tilde{p}^m$), and the mapped target (ground truth) patch-level representation coming from the Teacher

encoder ($\tilde{g}^m$);

$$L_{FP} = D(\tilde{p}, \tilde{g}) = \frac{1}{N_m} \sum_{i=1}^{N_m} \|(\tilde{p}_i - \tilde{g}_i)\|_1 \tag{4.8}$$

Overall Objective Function. The aim of FILS is to learn semantically meaningful video representations by jointly optimising video–text contrastive alignment and feature prediction in a shared semantic space. Therefore, the final objective of FILS is a combination of video-text contrastive loss and feature prediction loss. λ_1 and λ_2 adjust the weighting between our proposed contrastive loss(ActCLIP) and feature prediction loss(FP).

$$L_{FILS} = \lambda_1 L_{ActCLIP} + \lambda_2 L_{FP} \tag{4.9}$$

4.2 Datasets and Metrics

To illustrate the performance of our approach, we apply the proposed method to four datasets: Something-Something V2 (SSV2) [87], EPIC-KITCHENS-100 (EK100) [47], Charades-Ego [222], and Extended GTEA Gaze+ (EGTEA) [144]. The first two datasets (SSV2 and EK100) were already introduced in Chapter 3 and serve as our primary benchmarks for large-scale training and evaluation. Here, we briefly restate their key characteristics before introducing the smaller egocentric datasets used to assess transferability and robustness.

Something-Something V2 (SSV2) As detailed in Chapter 3, SSV2 comprises 220,847 short video clips covering 174 action categories, depicting everyday human–object interactions captured from a third-person perspective. It is particularly challenging for action recognition because many videos involve the same objects and human hands performing very similar motions that belong to different action categories, making it difficult for models to distinguish actions based solely on temporal and motion cues.

EPIC-KITCHENS-100 (EK100) is the largest egocentric action recognition dataset, containing 700 videos (over 100 hours) divided into short action segments with verb–noun annotations. It covers 97 verbs, 300 nouns, and 3,806 actions. The dataset is used to benchmark first-person understanding under realistic challenges such as occlusions, limited field of view, and camera motion.

Charades-Ego is a dataset comprising 7,860 first- and third-person recordings of everyday indoor activities across 157 action classes, totalling 68.8 hours of video. There are 34.4 hours of first-person and 34.4 hours of third-person recordings, with an average of 8.72 activities (68,536 activity instances) in each video. The videos are from 112 different rooms all over the world. In our experiments, we use the first-person subset, consisting of 3,085 videos for training and 846 for testing purposes. Although smaller in scale, its egocentric videos offer greater variation in subjects and scenes compared to EK100, making it a useful benchmark for evaluating first-person action recognition under controlled yet visually diverse conditions.

Extended GTEA Gaze+(EGTEA) This dataset provides 29 hours of egocentric cooking activities recorded from 32 participants performing seven meal preparation tasks in a realistic kitchen environment. It includes 86 video sessions with 10,321 annotated action segments across 106 classes, with an average segment length of 3.2 seconds. Its strong focus on hand–object interactions and gaze behaviour makes it ideal for validating the fine-grained recognition capabilities of models trained on larger egocentric datasets.

Together, these four datasets enable a comprehensive evaluation across third-person and first-person settings, from large-scale, diverse benchmarks to smaller, task-specific datasets. Fig. 4.3 provides visual examples of these datasets to highlight their differences in viewpoint and scene context.

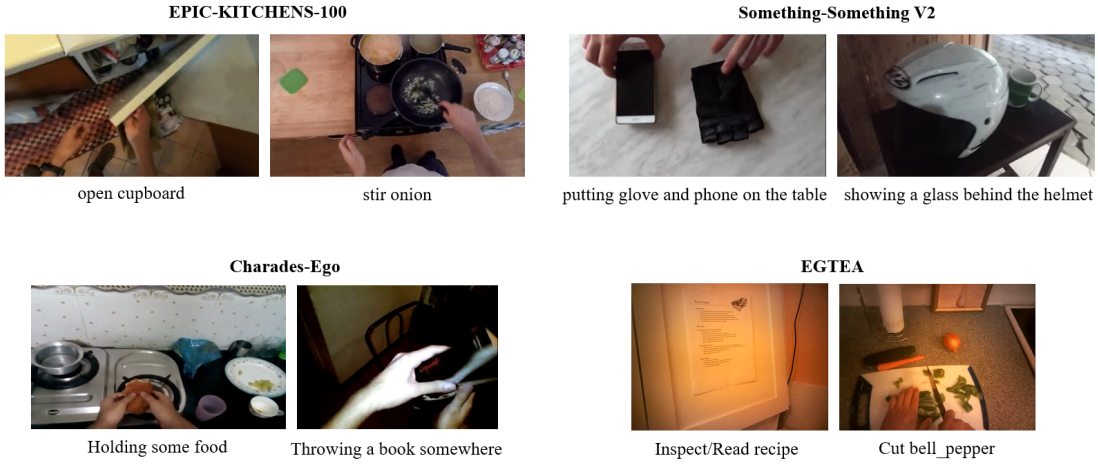


Figure 4.3: Sample frames from the datasets used in this chapter. The examples highlight differences in viewpoint, scene context, and interaction style. Each sample is annotated with its corresponding action class.

4.3 Implementation Details

For self-supervision, we sample 16 RGB frames from each video as a clip with a dynamic stride (depending on the number of raw video frames). The frame resolution is 224×224, and random resized cropping is used as an augmentation. Our spatial patch size is 16×16 with a temporal patch size of 2. Therefore, each clip is split into non-overlapping 8×14×14 tubes, yielding 1568 tokens. The encoder in each of our experiments is the ViT-B/16 architecture [59], trained on the EPIC-KITCHENS-100 (EK100) and Something-Something V2 (SSV2) datasets. We use the AdamW optimiser [169] with $(\beta_1, \beta_2) = (0.9,0.95)$ and a weight decay of 0.05. The learning rate starts at 1e-6, grows linearly to a peak of 1.5e-4 in the first epoch, and then uses a half-wave cosine schedule to decline to 1e-5 progressively. λ_1 and λ_2, which adjust the weighting between our proposed contrastive loss (ActCLIP) and feature prediction loss (FP), are set to 1. The predictor has six additional transformer layers. As a preprocessing step, we generate synthetic captions for all videos in the training set using a video-to-text model (VideoBLIP) [302]. Our text encoder employs a CLIP-based encoder with frozen weights from ViFi-CLIP [206]. The pretraining is conducted for 800 epochs, and following [321], we employ flash attention [48] to lessen the memory bottleneck of attention operations; it is more efficient than standard attention techniques in terms of IO complexity. We also leverage gradient checkpointing [31, 39] for training the transformer to reduce memory cost and use PyTorch gradient checkpointing. For ViT-B, we employ a batch size of 64 per GPU across 4 GPUs, for a total batch size of 256, significantly smaller than the thousands used in previous studies.

We fix the parameters for all our finetuning experiments using the same hyperparameters as the trained baselines. We use AdamW with a momentum of 0.9 and a weight decay of 0.05 to finetune the pretrained model on EK100 and SSV2 for a specific number of epochs on EK100, Charades-Ego, EGTEA, and SSV2. We employ cosine annealing with a warm-up, where the base learning rate starts at 1e-6, increases linearly to a peak of 1.5e-3 in the first epoch, and then gradually decreases to 1e-6 using a half-wave cosine schedule. We replace the linear projection head for action classification with a dataset-specific dimension head. We marginalise the action-level probability to obtain verb- and noun-level accuracy. We also employ 0.8 mixup and 0.1 label smoothing. We input 16 sampled frames for each video clip during training and testing, and resized the shorter side to 256 pixels. Next, we take a 224x224 crop and apply data

augmentation using conventional Random Resized Crop (0.08,1.0) and a Horizontal Flip (0.5), fused at the video-decoding side. We take the centre 224×224 crop at inference and scale the shorter side to 224 pixels. For ViT-B, we employ a batch size of 64 per GPU over 4 GPUs.

The computational complexity of FILS is reported in terms of floating-point operations (GFLOPs). For a 16-frame input clip at a resolution of 224×224, the full FILS model requires 23.68 GFLOPs at inference. Most of the computation arises from the visual encoder and the video-to-text mapping modules, while the textual encoder contributes a smaller portion of the overall cost. This demonstrates that FILS achieves enhanced semantic modelling with a moderate computational footprint.

4.4 Results

We assess the quality of the representations learnt by FILS on the challenging action recognition task, which is difficult to automate despite being easy for humans. Our experiments show that FILS achieves superior performance across the metrics on the utilised action recognition dataset, demonstrating how vision self-supervision enhances the vision-language contrastive approach.

4.4.1 Action Recognition Task

We assess the learnt video representation by finetuning the video encoder for action classification. In line with previous studies [238, 322], after finetuning the video encoder, top-1 accuracy is reported on verbs, nouns, and actions for EPIC-KITCHENS-100 [47] and actions for Something-Something V2 [87]. For EGTEA [144], in addition to top-1 accuracy, we include mean class accuracy, utilising the initial train/test split, and for Charades-Ego [222], the evaluation metric is mean average precision (mAP). The recognition accuracies for the EK100 and SSV2 datasets are reported in table 4.1 and 4.2, respectively. Our proposed method notably enhances the effectiveness of the existing supervised and self-supervised techniques for action recognition over ViT-B. We have at least 1.9% improvement on action, verb, and noun on EK100 over the best results in the literature, AVION [321] and LaViLa [322]. Similarly, on the SSV2 action, the performance increase is $+0.9\%$ compared to the highest accuracy in the literature using ViT-B. To further evaluate its transferability, we use the FILS model pretrained on EK100

Table 4.1: Performance of action recognition on EK100. FILS outperforms all prior works regarding action-level top-1 accuracy. In the table below, *p-data* and *L* mention pretraining data utilised for the incorporation of language during training, respectively. Models marked with * indicate results reproduced by us using the official code released by the authors.

Method	Backbone	p-data	L	Verb Top-1	Noun Top-1	Action Top-1
SlowFast [68]	ResNet101	K400	×	65.6	50.0	38.5
TSM [149]	ResNet50	IN-1K	×	67.9	49.0	38.3
Mformer [192]	ViT-L	IN-21K+K400	×	67.1	57.6	44.1
Video Swin [167]	Swin-B	K400	×	67.8	57.0	46.1
ViViT FE [6]	ViT-L	IN-21k+K400	×	66.4	56.8	44.0
IPL [263]	I3D	K400	✓	68.6	51.2	41.0
Omnivore [84]	Swin-B	IN+K400+SUN	×	69.5	61.7	49.9
MeMViT [271]	ViT-B	K600	×	71.4	60.3	48.4
MTV [290]	MTV-B	WTS-60M	×	69.9	63.9	50.5
LaViLa [322]	TSF-B	WIT+Ego4D	✓	69.0	58.4	46.9
AVION [321]	ViT-B	WIT+Ego4D	✓	70.0	59.8	49.1
VideoMAE* [238]	ViT-B	EK100	✓	-	-	48.5
FILS(ours)	ViT-B	EK100	✓	72.2	61.7	51.0

and finetuned on the Charades-Ego and EGTEA datasets, which are visually distinct from EK100. FILS improves mAP on Charades-Ego by $+1.3\%$ and top-1 accuracy on EGTEA by $+1.03\%$ compared to the baseline LAVILA model [322] after finetuning. Table 4.3 and table 4.4 demonstrate our superior performance compared to the current state-of-the-art on these datasets. Therefore, instead of relying on large-scale datasets like the Kinetics [118, 24], which include URL links to up to 650,000 video clips, our method improved performance even when trained on an approximately 10x smaller dataset.

4.4.2 Ablation Study

Self-supervision Strategy

The core insight of FILS is to perform masked feature prediction in language semantic space. In Sec. 4.1.2, we introduce our initial objective, FP, which entails predicting features. To demonstrate the benefits of predicting patch features instead of reconstructing missing patches in

Table 4.2: Something-Something V2 Action Recognition. On SSV2, FILS consistently outperforms previous approaches with higher action-level top-1 accuracy. The table specifies *p-data*, denoting the pretraining dataset, and *L*, indicating whether language was incorporated during training. Results marked with * correspond to our reproduction using the official implementation provided by the authors.

Method	Backbone	p-data	L	Top-1
SlowFast [68]	ResNet101	K400	×	63.1
TSM [149]	ResNet50	K400	×	63.4
TimeSformer [17]	ViT-L	IN-21K	×	62.4
Mformer [192]	ViT-L	IN-21K+K400	×	68.1
Video Swin [167]	Swin-B	K400	×	69.6
ViViT FE [6]	ViT-L	IN-21k+K400	×	65.9
VIMPAC [234]	ViT-L	HowTo100M	✓	68.1
BEVT [259]	Swin-B	IN-1K + K400	×	70.6
VideoMAE [238]	ViT-B	SSV2	×	70.8
OmniMAE [83]	ViT-B	IN-1K + SSv2	×	69.5
Omnivore [84]	Swin-B	IN-21k+K400	×	71.4
VideoMAE V2 [252]	ViT-B	UnlabeledHybrid	×	71.2
FILS(ours)	ViT-B	SSV2	✓	72.1

Table 4.3: Charades-Ego Action Recognition. FILS achieves substantial improvements on Charades-Ego, outperforming prior work, even though Charades-Ego videos differ visually from the EK100 and SSV2, datasets used for FILS pretraining. The table reports *p-data* and *L*, referring to pretraining data and language incorporation during training, respectively.

Method	Backbone	p-data	L	mAP
ActorObserverNet [221]	ResNet-152	Charades	×	20
SSDA [40]	I3D	Charades-Ego	×	25.8
Ego-Exo [141]	SlowFast-R101	Kinetics-400	×	30.1
EgoVLP [151]	TSF-B	Ego4D	✓	32.1
HierVL-Avg [7]	ViT-Base	Ego4D	✓	32.6
HierVL-SA [7]	ViT-Base	Ego4D	✓	33.8
EgoVLPv2 [195]	TSF-B	EgoClip	✓	34.1
LaViLA [322]	TSF-B	WIT+Ego4D	✓	33.7
FILS(ours)	ViT-Base	EK100	✓	**34.4**
FILS(ours)	ViT-Base	SSV2	✓	**34.2**

the pixel domain using mean square error loss, we train a pair of ViT-B/16 models on the EK100 dataset using feature prediction loss and mean squared error loss without the contrastive language contribution. To evaluate the action recognition accuracy of models trained with FP and MSE objectives, we finetuned a pretrained model on EK100. The model trained with the FP objective exhibits superior action accuracy, 50.3%, compared to the model trained with MSE,

Table 4.4: EGTEA Action Recognition. FILS shows clear performance gains on EGTEA, surpassing existing methods, despite the visual domain gap between EGTEA and the pretraining datasets EK100 and SSV2. The table lists *p-data*, the pretraining dataset, and *L*, language incorporation during training. Results marked with * denote models trained by us using the authors' provided code.

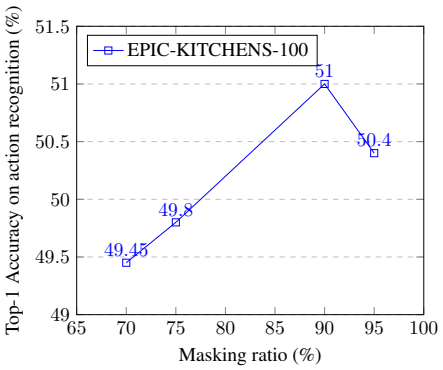
Method	Backbone	p-data	L	Top-1 Acc.	Mean Acc.
Li et al. [144]	I3D	K400	×	-	53.30
LSTA [227]	ConvLSTM	IN-1k	×	61.86	53.00
IPL [263]	I3D	K400	✓	-	60.15
MTCN [119]	SlowFast	K400+VGG-S	✓	73.59	65.87
LaViLA [322]	TSF-B	WIT+Ego4D	✓	77.45	70.12
FILS(ours)	ViT-Base	EK100	✓	**78.48**	**71.20**
FILS(ours)	ViT-Base	SSV2	✓	**78.57**	**71.31**

Table 4.5: Ablation study on two patch-wise contrastive scenarios and masking ratio.

(a) ActCLIP strategies

strategy	number of iteration	Action Top-1 Acc.
patch	10	38.0
patch	30	43.5
patch	50	46.3
patch-average	1	51.0

(b) Masking ratio



48.5%. Both are lower than the 51.0% achieved on EK100 action recognition via our full FILS models, evidence of the effectiveness of our proposed method, which involves feature prediction within the language space.

ActCLIP strategies

In our patch-wise ActCLIP framework, we applied contrastive learning between the average feature extracted from patches within the action area and text features. To examine the impact of the number of patches used, we conducted an ablation study (table 4.5(a)). This involved randomly selecting one patch within the action area (patch strategy) instead of averaging across all patches within the action area (patch-average strategy). We repeated this process several times

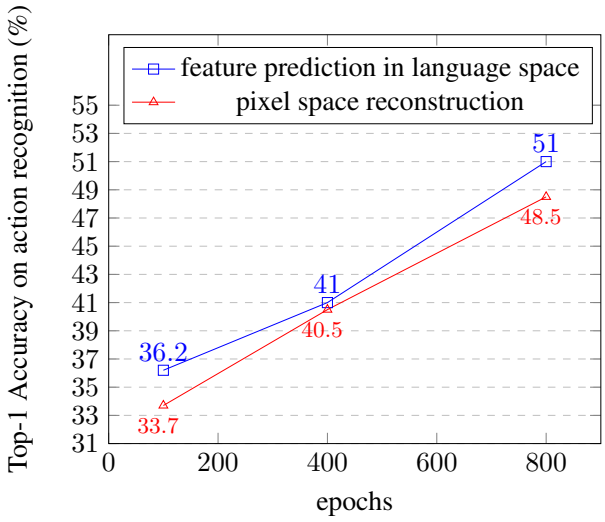


Figure 4.4: The impact of varying pretraining epochs on the EPIC-KITCHENS-100 dataset. There is a consistent upward trend in action recognition accuracy with an increase in the number of pretraining epochs.

and documented the results alongside the number of iterations in table 4.5(a). As anticipated, increasing the number of iterations improved performance, approaching the patch-average strategy utilised in ActCLIP.

Masking Ratio

We compare different masking ratios in table 4.5(b) on our proposed method using the EPIC-KITCHENS-100 dataset. Increasing the masking ratio from 70% to 95% for tube masking, we find that the performance is higher with an extremely high ratio of 90%.

Comparison FILS with Pixel Reconstruction

In fig. 4.4, we demonstrate that our proposed method (FILS) exhibits notable enhancements, surpassing the performance of the pixel space reconstruction technique [238, 321], which attained its results relying on extensive epochs of pretraining. We present the action recognition results of FILS and the pixel-reconstruction baseline on the EPIC-KITCHENS dataset using a ViT-B/16 pretrained on the self-supervised objective for 100, 400, and 800 epochs. This experiment aims to demonstrate our results' consistency and superior FILS performance over the pixel space reconstruction method, even with fewer training epochs.

4.4.3 Attention Visualization

To gain deeper insight into the learnt representations by FILS, we employ Grad-CAMs [212] to visualise the prominent areas that significantly contribute to the accomplishment of the action recognition task. This visualisation helps us better comprehend the spatiotemporal cues acquired during the self-supervised learning step. In fig. 4.5, we illustrate attention visualisation for a few sample videos selected from the EPIC-KITCHENS-100 dataset using the models trained with different training strategies: our proposed FILS, our first objective, which is FP, and MSE in the pixel domain. Sec. 4.4.2 discovers the comparison among these strategies. We selected instances that FILS correctly classified, whereas feature prediction (FP) and mean squared error (MSE) failed to do so. Our visualisation study shows that, compared to ViTs trained with MSE and FP objectives, the one trained with FILS generates attention heatmaps that emphasise the area where the action happens and are more effective at classifying the videos' actions. The attention maps demonstrate how well our proposed self-supervised technique (FILS) represents the potential semantic region in the video and acquires an understanding of spatiotemporal relationships by linking pertinent areas, distinctly revealing the semantics of actions. These findings further demonstrate the effectiveness of conducting contrastive learning between the patches within the action area and the relevant text in our proposed self-supervised technique (FILS).

In fig. 4.6, we display the Grad-CAM for the challenging examples of EPIC-KITCHENS-100, Something-Something V2, and EGTEA. For each video, we show attention heatmaps on the first, middle, and last frames to highlight how the models capture temporal dynamics.

4.4.4 FILS learns semantic representations

Our proposed approach predicts visual features within a language space constructed through contrastive learning between patches in the identified action areas and their corresponding text, enabling the model to receive implicit guidance from the textual context. Using videos from the EK100, we calculate the similarity of features between video patches and their corresponding text features, which is the action label. Fig. 4.7 compares our proposed contrastive strategy (first objective – ActCLIP) and our FILS. FILS significantly enhances the contrastive objective between language and vision. This is evident in the heatmap's improved localisation and reduced

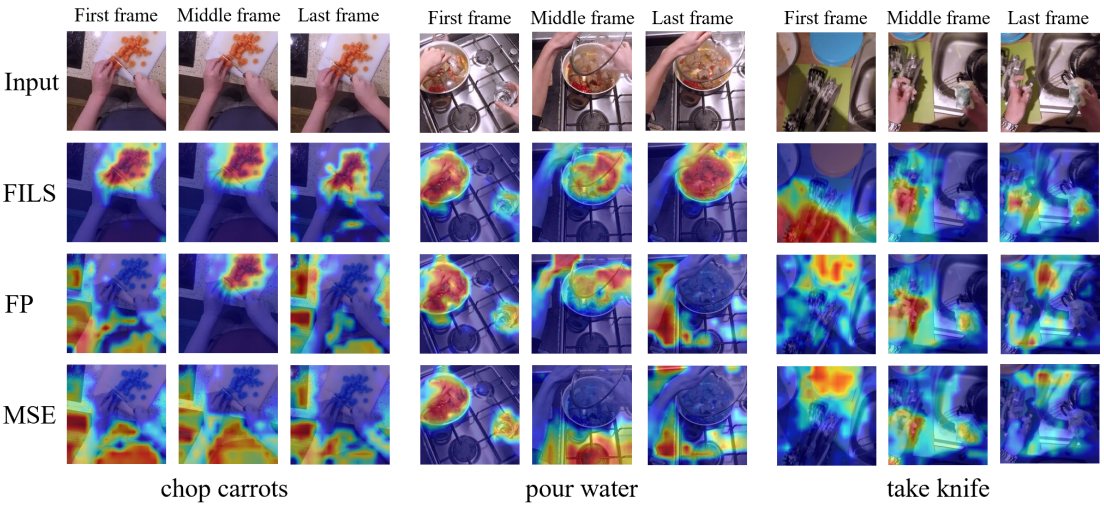


Figure 4.5: Attention heatmaps were generated for the initial, central, and final frames of the EK100 using the last transformer layer of the model trained with self-supervised strategies, including FILS, our second objective (FP), and pixel-domain reconstruction (MSE) after masking.

noise, which now exhibits greater concentration around the object and action regions. To improve visualisation, we smoothed the patch-wise attention blocks with Gaussian blurring.

To further analyse the semantic representations learned by FILS, we visualise patch-level similarity maps using verb-only and noun-only text embeddings. As shown in Fig. 4.8, FILS produces more distinct and localised activations, with verbs emphasising action dynamics and nouns focusing on object regions. Compared to ActCLIP, the activations are less noisy and better aligned with semantic roles, indicating improved disentanglement of action and object information.

4.4.5 Synthetic Captions

In our self-supervision step, as a preliminary step, we use a video-to-text model (VideoBLIP) [302] to generate synthetic captions for all of the videos in the training set. This process was conducted for both the EPIC-KITCHENS-100 and Something-Something V2 datasets. In fig. 4.9, we present examples of generated captions alongside their respective labels for the training sets of these two datasets. These synthetic captions for videos have demonstrated remarkable effectiveness, providing detailed scene descriptions and capturing concepts closely related to video labels, which could enhance the interpretability of video content. Using these generated captions in our training process could enhance the comprehension and performance of our

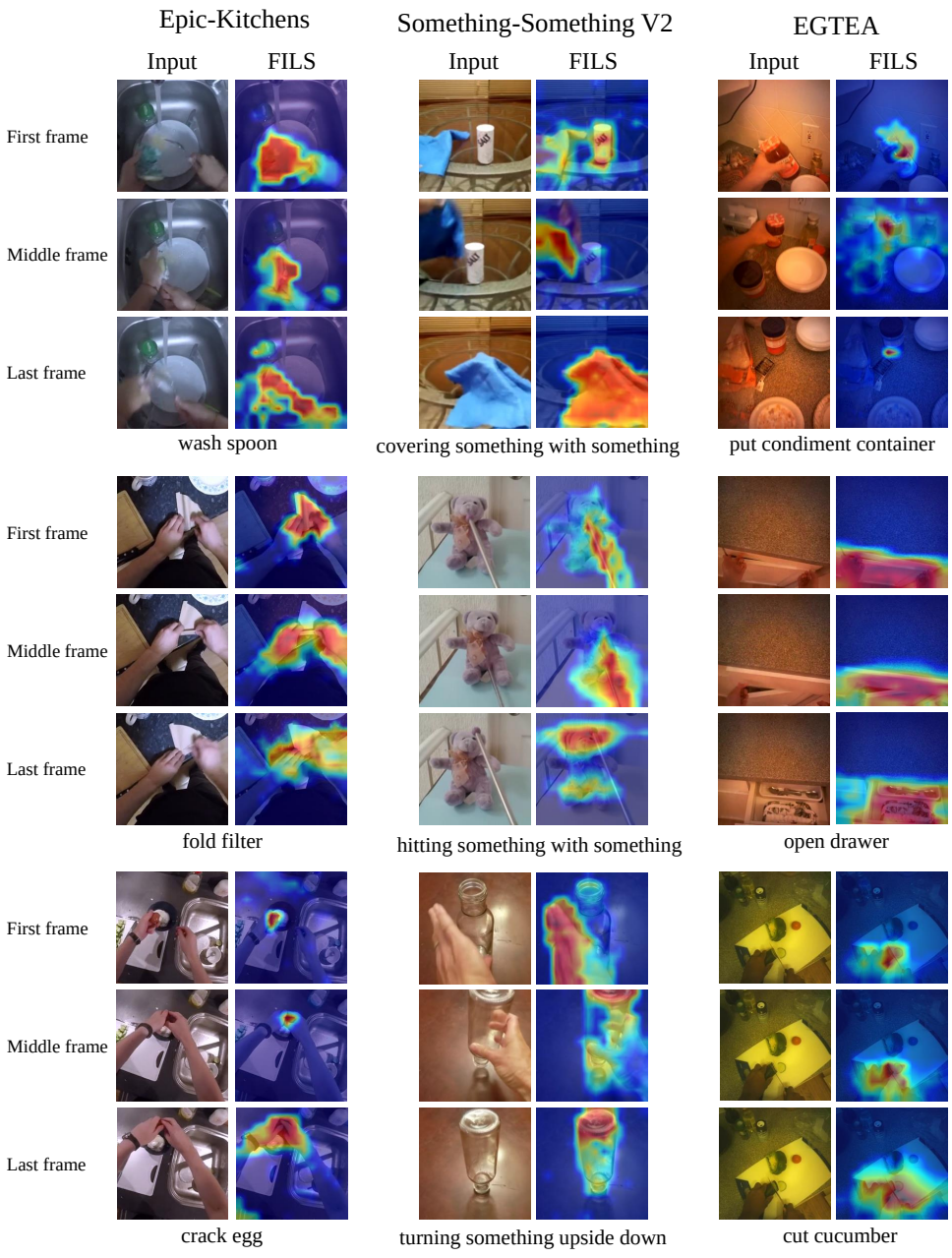


Figure 4.6: Visualisation by Grad-CAM on EPIC-KITCHENS-100, Something-Something V2 and EGTEA.

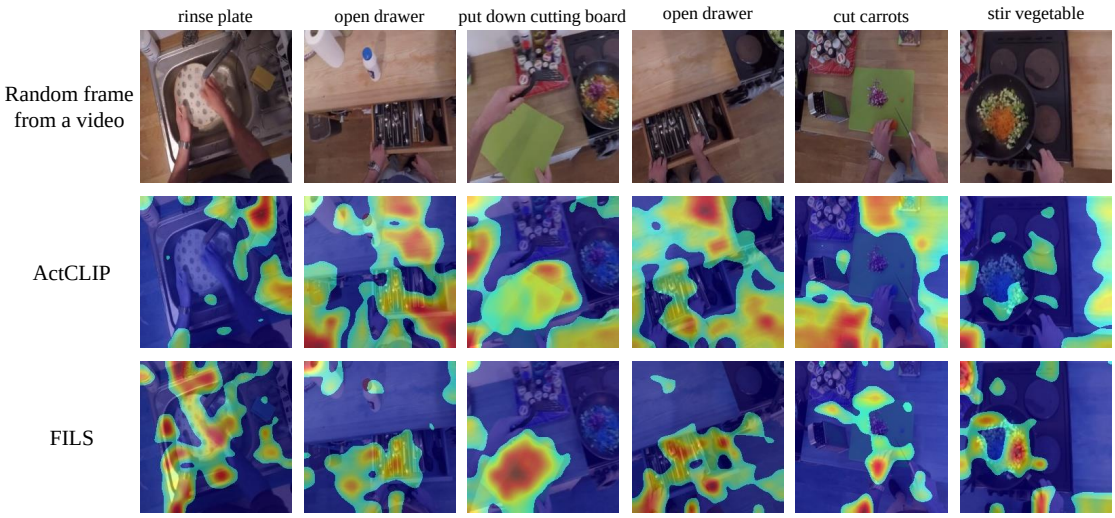


Figure 4.7: visualisation of the similarity between text and video features for the EK100 dataset. The provided text is the video's action label.

model, as these enriched input features are semantically meaningful, which is crucial for training models.

4.5 Conclusion

The FILS framework represents a significant advancement in self-supervised video representation learning, addressing critical challenges in understanding complex video data. By combining masked feature prediction with semantic language guidance and motion-aware contrastive learning, FILS offers a novel approach to learning rich and transferable video representations. Unlike traditional methods that struggle to capture temporal dynamics or rely heavily on labelled data, FILS leverages text captions and motion saliency to align visual features with meaningful semantics. This integration bridges the gap between raw video data and high-level activity comprehension, making it a powerful tool for video understanding tasks.

One of the key contributions of FILS is its ability to focus on action-relevant regions within videos, ensuring that the learnt representations prioritise areas most critical for understanding human activities. The use of the motion map technique to identify motion saliency, combined with ActCLIP's contrastive learning between patches and text captions, enhances the model's ability to capture fine-grained temporal dependencies and semantic richness. Additionally, FILS

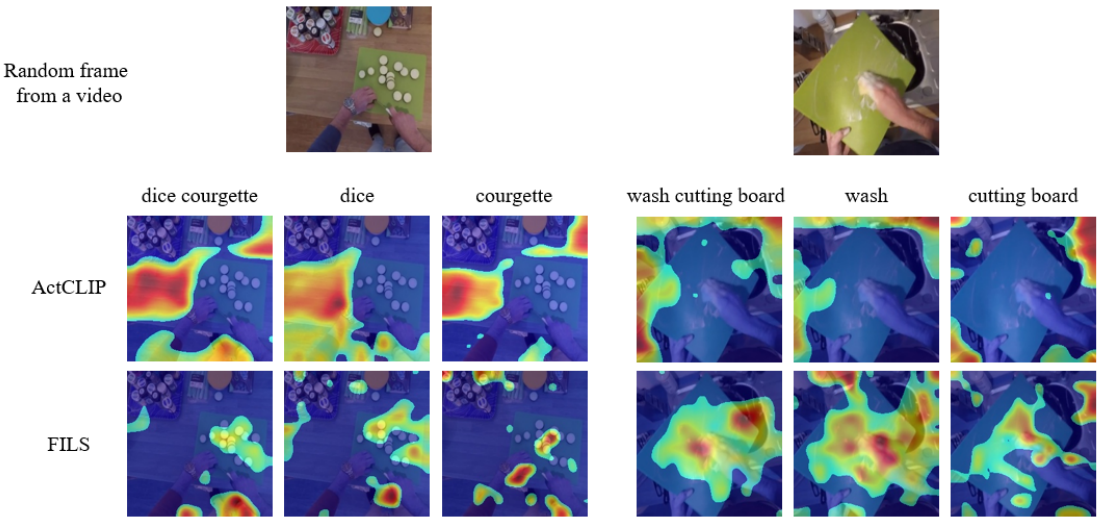


Figure 4.8: Verb- and noun-level language–vision alignment on EPIC-KITCHENS-100. Patch-wise similarity heatmaps are computed using verb and noun text embeddings separately. FILS yields clearer and more semantically focused activations than ActCLIP, highlighting action-related regions for verbs and object-centric regions for nouns.

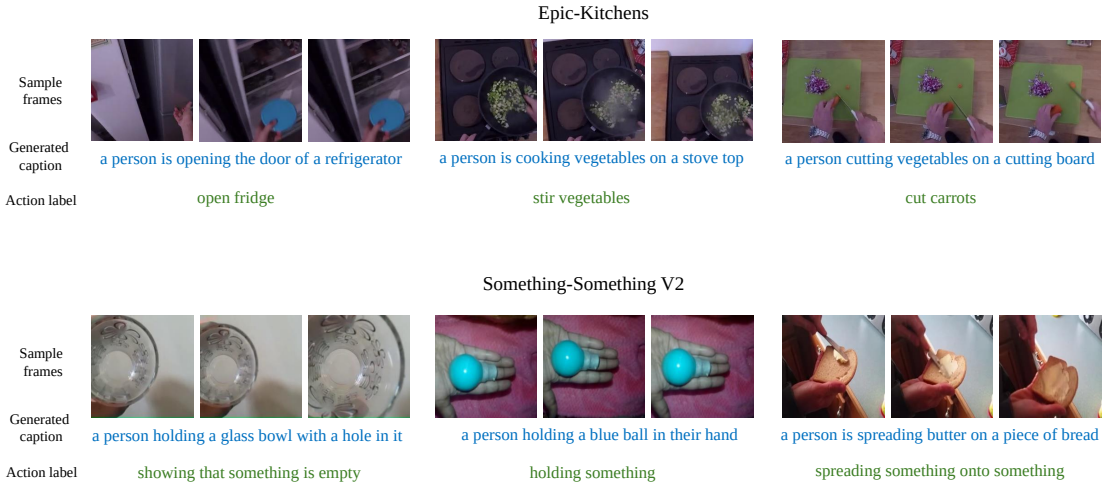


Figure 4.9: Synthetic captions for some instances from the training set of EPIC-KITCHENS-100 and Something-Something v2. VideoBLIP often captures good spatial and temporal details.

avoids pixel-level reconstruction, making it computationally efficient while maintaining state-of-the-art performance across diverse datasets. Experimental evaluations on benchmarks such as EPIC-KITCHENS-100, Something-SomethingV2, Charades-Ego, and EGTEA demonstrate the robustness and transferability of FILS across domains, highlighting its potential for real-world applications.

The success of FILS has broader implications for multimodal learning and video analysis. By treating text as a supervisory signal, FILS demonstrates how language can serve as a guiding framework for structuring visual semantics. This approach reduces reliance on labelled data while improving interpretability, an essential feature for applications requiring explainability, such as healthcare monitoring or assistive robotics. Furthermore, its motion-guided contrastive learning offers valuable insights into handling video-specific challenges like temporal redundancy and action localisation, paving the way for advancements in related tasks such as video summarisation or anomaly detection. While FILS achieves impressive results, there are opportunities for future work to further enhance its capabilities. Incorporating audio cues could enrich multimodal alignment, enabling better synchronisation of visual and auditory events.

In conclusion, FILS introduces a novel paradigm for self-supervised video representation learning by unifying semantic language guidance, motion saliency, and efficient masked feature prediction. Its ability to focus on action-relevant regions while leveraging textual information makes it a versatile and impactful framework for video understanding tasks. As video data continues to dominate digital ecosystems, approaches like FILS will play a crucial role in transforming raw visual streams into actionable knowledge, offering a strong foundation for future innovations in multimodal AI systems. In the next chapter, DEL extends these ideas to the more complex challenge of dense multimodal event localisation and classification of multiple overlapping events in untrimmed videos. While FILS focused on bridging video and text to enhance representation learning, DEL broadens this perspective by integrating audio alongside vision, addressing the problem of overlapping events and asynchronous multimodal signals.

Chapter 5

DEL: Dense Event Localisation for Multimodal Audio-Visual Understanding

The growing demand for intelligent video understanding systems has brought Temporal Action Localisation (TAL) to the forefront of computer vision research. It is worth distinguishing TAL from the related task of video temporal grounding (also known as video moment retrieval). While TAL aims to automatically detect, localise, and classify all action instances in an untrimmed video without any external query, video temporal grounding is a query-driven task that retrieves a single temporal segment corresponding to a given semantic description, typically provided in natural language. As a result, TAL emphasises dense temporal detection and boundary precision, making it more suitable for discovering overlapping and asynchronous events in untrimmed videos. Following the exploration of language-guided video representation learning in the previous chapter, this chapter extends the investigation of multimodal understanding to the audio-visual domain. While FILS focused on aligning visual and linguistic modalities to enhance semantic representation learning, the proposed DEL framework incorporates auditory cues for dense event localisation in untrimmed videos, effectively addressing challenges such as overlapping events and temporal asynchrony. Although many recent approaches explore vision–language representations, DEL deliberately focuses on audio–visual multimodality. Dense event localisation relies on precise temporal cues and event boundaries, which are directly

encoded in audio and visual streams, whereas language typically provides higher-level semantic information that is weakly aligned with fine-grained temporal structure. Moreover, existing TAL benchmarks provide consistent supervision only for audio and visual modalities, making tri-modal learning impractical for this task.

TAL is a crucial and challenging task in video understanding. It focuses on identifying the temporal boundaries of action instances and classifying them within untrimmed videos, providing critical insights into real-world video content [240]. Understanding real-world scenes and events is inherently a multimodal perception process, integrating visual and auditory cues to achieve comprehensive video understanding [116, 184, 190]. This task becomes increasingly challenging in scenarios involving overlapping events, asynchronous modalities, and complex temporal dependencies. For instance, distinguishing between a person speaking and merely mouthing words is difficult with visual information alone, as both actions involve similar lip movements. Incorporating audio cues, however, helps resolve this ambiguity by detecting the presence or absence of vocal sounds and has proven to be a promising direction. Despite this potential, existing multimodal approaches often fail due to misaligned features, inadequate temporal modelling, and limited capacity to handle concurrent and densely overlapping events. While visual and auditory information are inherently complementary, effectively modelling multimodal interactions remains a significant challenge due to factors such as modality misalignment, varying action durations, and the intricate dependencies across modalities, particularly in scenarios involving overlapping or concurrent events [237].

To illustrate the inherent challenges in dense audio-visual event localisation, consider real-world scenarios in which multiple events overlap or occur asynchronously, each with varying duration and intensity. As shown in fig. 5.1, unimodal approaches that rely solely on either visual or auditory cues often struggle to accurately separate concurrent activities, leading to incomplete or ambiguous detections. For instance, a visual-only model may misinterpret subtle actions that depend on sound, while an audio-only model may fail to distinguish between similar background noises without visual context. These observations highlight the need for a unified framework that can capture fine-grained temporal dependencies and align cross-modal information. DEL is designed to address this challenge by jointly modelling audio and visual interactions, enabling precise localisation of both short- and long-duration events even when they co-occur.

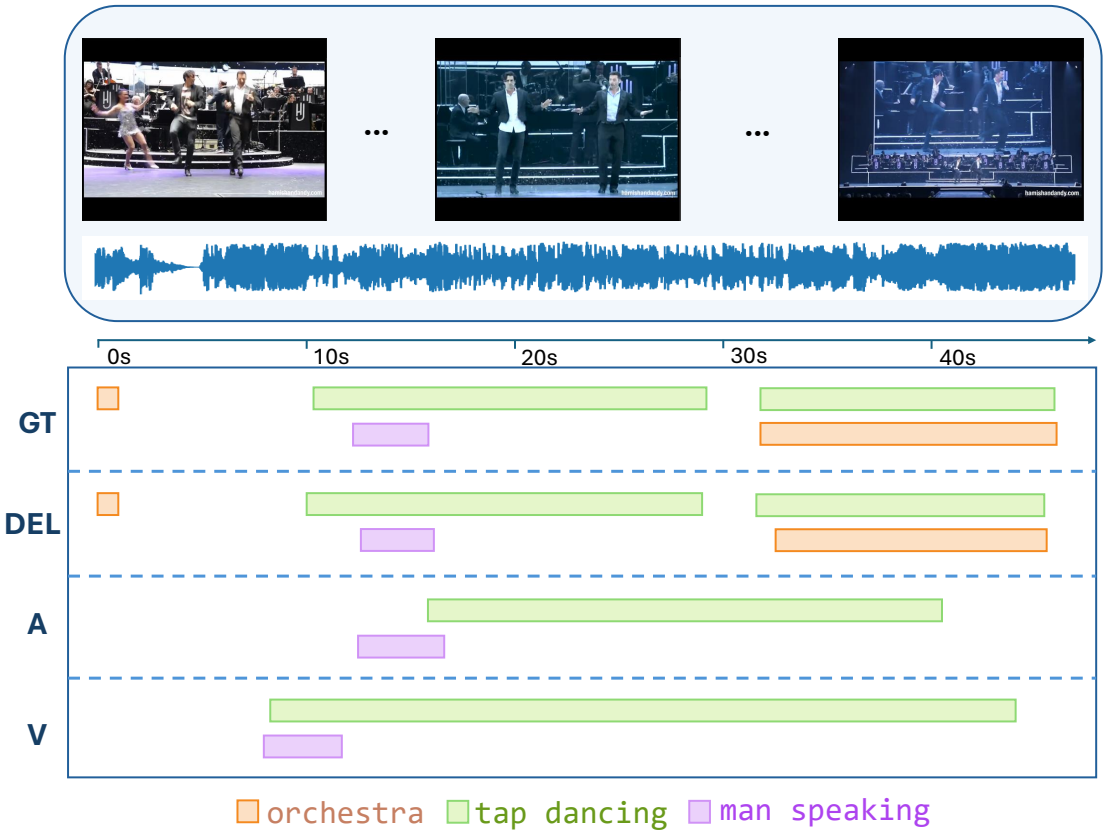


Figure 5.1: Real-world videos contain overlapping events of varying durations, making precise localisation challenging. The image presents ground-truth (GT) annotations alongside the predictions of DEL, an audio-only model (A), and a visual-only model (V). The unimodal models struggle with a certain category. In contrast, **DEL** effectively detects and classifies both short- and long-duration events, even when they co-occur, demonstrating its superior multimodal fusion capability.

Prior efforts have made progress in Audio-Visual Event Localisation (AVEL) for trimmed videos, where each video typically contains a single, isolated audio-visual event [80, 244, 196]. Unlike AVEL, dense localisation of audio-visual events requires identifying and recognising all events within an untrimmed video, capturing short- and long-duration interactions that may overlap [81, 112, 21, 47]. Recent advances have demonstrated the effectiveness of transformer networks and Feature Pyramid Networks (FPN) for TAL [291, 311, 269, 218], significantly improving performance by leveraging multi-resolution visual features. However, those approaches overlook the role of audio.

A persistent challenge in audio-visual event localisation is the effective extraction and fusion of information across modalities, especially when multiple events co-occur. Traditional approaches

often treat audio and visual modalities separately or rely on late fusion strategies [30, 191, 210], limiting their ability to capture fine-grained temporal dependencies. Moreover, most multimodal methods employ pretrained models to extract features, which, while beneficial for representation learning, often lead to misaligned features due to differences in pretraining objectives and domain shifts between audio and visual modalities. Additionally, contrastive learning techniques often focus on instance-level alignment while neglecting intra-video relationships, such as temporal coherence and cross-event correlations, which are crucial for distinguishing similar events that occur at different times. While audio-visual data provides rich contextual cues, effectively leveraging these multimodal representations for improved video interpretation remains an open problem.

Recent progress in video–language localisation has introduced techniques that address challenges closely related to those in audio–visual event localisation, particularly in terms of temporal boundary ambiguity and cross-modal alignment. For example, Huang et al. propose Elastic Moment Bounding to explicitly model uncertainty in temporal boundaries when localising activities in video [108], and also introduce cross-sentence temporal and semantic relation modelling to enforce consistency across multiple textual descriptions of video events [109]. Although these approaches are developed in the context of vision–text localisation, they provide important conceptual insights for multimodal temporal localisation more broadly. In the audio–visual setting considered in this work, however, the two modalities exhibit fundamentally different characteristics. Unlike text, which provides explicit semantic structure and relatively precise temporal cues, audio signals are continuous, noisy, and often weakly synchronised with visual content. As a result, techniques such as elastic boundary modelling and cross-sentence reasoning cannot be directly transferred to audio–visual event localisation without substantial modification.

To tackle the challenges of dense audio-visual action localisation, we propose DEL, a novel framework composed of three key modules. First, a **multimodal adaptive attention mechanism** aligns audio and visual cues through temporally adaptive interactions, ensuring temporal coherence and intra-modal consistency.

Second, at the core of our framework, we introduce a **score-based contrastive learning strategy** as the key innovation of our approach, which departs from conventional instance-level methods that rely on heuristic sampling. By incorporating token-level supervision with event-aware score

functions and category predictions, this module dynamically identifies meaningful positives and challenging hard negatives within a video, yielding temporally aligned, event-specific contrastive pairs. This approach explicitly addresses the limitations of existing methods that often focus on instance-level alignment while neglecting intra-video relationships, such as temporal coherence and cross-event correlations. By combining both, DEL ensures: (i) Enhanced feature discrimination within a single video. (ii) Improved cross-modal coherence and temporal alignment across multiple audio-visual events. (iii) Robust event localisation even in densely overlapping scenarios. Instead of manually defining positive and negative samples, our method introduces a feature scoring mechanism that enables the model to automatically assign confidence scores and select contrastive pairs during training. This enhances feature discrimination and cross-modal coherence for precise event localisation.

Finally, a **path aggregation network** fuses multiscale audio-visual features, preserving fine-grained temporal details while aggregating high-level semantics. Here, multiscale refers to modelling events at different temporal resolutions through feature-level temporal downsampling, enabling the localisation of both short- and long-duration events within the same video. Together, these three modules enable DEL to explicitly model cross-modal dependencies, enhance feature discrimination, and achieve robust localisation even in scenarios with densely overlapping events.

We consistently demonstrate superior performance and notable improvements in mean Average Precision (mAP) scores across multiple benchmarks, including UnAV-100, THUMOS14, ActivityNet 1.3, and EPIC-KITCHENS-100, using the same encoders as prior works. These results highlight the framework's ability to handle diverse scenarios involving overlapping events and asynchronous modalities.

In summary, DEL represents a significant advancement in multimodal video understanding by addressing longstanding challenges in TAL with innovative solutions. The framework's ability to dynamically align audio-visual features, model fine-grained temporal dependencies, and enhance feature discrimination through contrastive learning makes it a powerful tool for dense event localisation. As video data continues to grow in complexity and scale, DEL provides a scalable solution that can be applied across various domains such as surveillance, video retrieval, sports analytics, and human activity recognition. By bridging the gap between multimodal

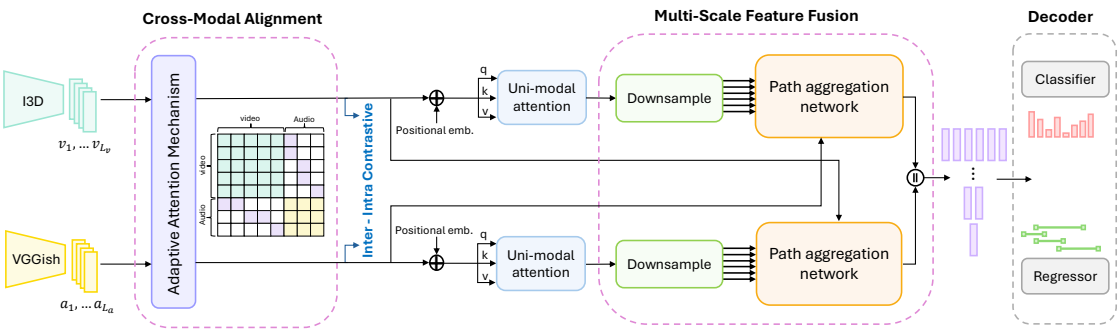


Figure 5.2: Overview of our proposed **DEL** framework. Our model integrates **(1)** an *adaptive attention mechanism* for aligning audio and visual features, **(2)** *inter- and intra-modal contrastive learning* to enhance event discrimination, and **(3)** a *multiscale path aggregation network* for feature fusion. $\|$ represents the concatenation operation.

representation learning and practical applications, this work paves the way for future research in understanding complex real-world video content.

5.1 Methodology

An overview of our framework is shown in fig. 5.2. Given untrimmed audio-visual input, the dense action localisation task aims to precisely detect the temporal boundaries of action instances, thereby requiring refinement of their semantic representations. To begin with, we tokenise the audio and visual data using pretrained feature extraction networks: a video backbone is used to encode each temporal segment into a visual token, and an audio encoder is used to extract the corresponding auditory token. Our method then aligns and fuses audio-visual features for precise event localisation through three key modules: (1) We employ an adaptive attention mechanism for cross-modal alignment, which dynamically aligns the audio and visual representations. (2) Next, score-based contrastive learning with both inter- and intra-modal objectives helps to dynamically select hard-negative samples to improve feature discrimination in the training process. (3) Finally, our path aggregation network for multiscale feature fusion facilitates multiscale feature extraction by generating multiple temporal representations and aggregating them to enhance temporal modelling.

We formulate dense audio-visual event localisation as a joint event classification and boundary estimation problem, enabling the automatic detection and temporal localisation of all events in untrimmed videos by jointly leveraging audio and visual cues. Given a video with audio-visual

data, we segment it into T paired segments, represented as

$$\mathbb{S} = \{(\mathbf{V_t}, \mathbf{A_t})\}_{t=1}^{T}, \tag{5.1}$$

where $\mathbf{V_t}$ and $\mathbf{A_t}$ denote the visual and audio features at time t. Note that the number of temporal segments, T, varies across videos. The ground-truth annotation is defined as

$$\mathcal{G} = \{g_n = (\tau_{start,n}, \tau_{end,n}, \lambda_n)\}_{n=1}^{N}, \tag{5.2}$$

where $\tau_{start,n}$ and $\tau_{end,n}$ indicate the event boundaries and $\lambda_n \in \Lambda$ represents the event class from a defined set of $|\Lambda|$ categories and N is the total number of events in the video. It is important to note that these annotations do not provide explicit pairing between audio and visual tokens. Instead, each event label is defined on the shared temporal axis of the video and applies jointly to both modalities, providing supervision at the event level rather than at the level of fine-grained audio–visual alignment.

During inference, our model predicts localised events

$$\hat{\mathbb{S}} = \{\hat{s}_t = (\delta_{start,t}, \delta_{end,t}, q(y_t))\}_{t=1}^{T} \tag{5.3}$$

where y_t denotes the ground-truth event label associated with segment t,and $q(y_t) \in [0,1]^{|\Lambda|}$ is the event classification probability. $\delta_{start,t}$ and $\delta_{end,t}$ represent estimated temporal offsets of the current time t for event boundaries. The final predictions are:

$$\begin{aligned} \hat{\lambda}_t &= \arg\max_{\lambda \in \Lambda} q(\lambda_t), \\ \hat{\tau}_{start,t} &= t - \delta_{start,t}, \\ \hat{\tau}_{end,t} &= t + \delta_{end,t}. \end{aligned} \tag{5.4}$$

This results in the fine-grained localisation and classification of overlapping audio-visual events.

5.1.1 Adaptive Attention for Cross-Modal Alignment

A key challenge in dense audio-visual event localisation is modality misalignment. Audio cues and visual evidence often occur with temporal shifts or are obscured by background noise. This

makes unconstrained cross-modal attention prone to spurious matches (e.g., background sounds attending to unrelated frames).

To address this, we propose an adaptive attention mechanism that dynamically adjusts feature importance to better synchronise audio and visual signals, enhancing cross-modal coherence. Central to this is an event-aligned attention mask $\mathbf{M}$, which guides feature interactions within and across modalities. Unlike standard self-attention, our approach introduces a learnt mask matrix that constrains attention based on temporal alignment, emphasising interactions likely to represent the same event while suppressing irrelevant ones. This design reduces noise, stabilises training, and improves interpretability by focusing attention on semantically meaningful correspondences. To compute this adaptive attention, an attention mask $\mathbf{M}$, is initialised with the sizing, $\mathbf{M} \in \mathbb{R}^{(L_v+L_a)\times(L_v+L_a)}$, where L_v and L_a denote the lengths of the video and audio features from the sequences, respectively.

Then, given a concatenated input $\mathbf{X} = [\mathbf{V}|\mathbf{A}] \in \mathbb{R}^{(L_v+L_a)\times d}$ where X contains both video and audio features, we construct the query ($\mathbf{Q}$) and key ($\mathbf{K}$) matrices via linear transformations:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V \tag{5.5}$$

$\mathbf{Q}$ encodes the information each token seeks, $\mathbf{K}$ represents how tokens signal their relevance for matching, and $\mathbf{V}$ carries the content that is aggregated through attention to enable effective audio–visual feature fusion. $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$ are learnable projection matrices. Next, we compute the adaptive attention as

$$\text{aat}_{i,j} = \frac{m_{i,j}\exp(\mathbf{Q}_i\mathbf{K}_j^T/\sqrt{d})}{\sum_k m_{i,k}\exp(\mathbf{Q}_i\mathbf{K}_k^T/\sqrt{d})} \tag{5.6}$$

where $m_{i,j}$ is the (i,j)-th entry of the attention mask matrix M and $i, j \in [1, L_v + L_a]$ represent the indices of the adaptive attention matrix, $\mathbf{AAT}$. d denotes the dimensionality of the token feature embeddings and k is the index over all candidates used to normalise the attention distribution. The mask $\mathbf{M}$ aims to emphasise and learn feature interactions corresponding to the same event across both intra- and inter-modality. For intra-modality attention, we apply a standard global attention mechanism, allowing features within the same modality to attend to each other by setting the corresponding entries in the mask to 1. For cross-modality attention,

only entries corresponding to the same temporal segment are assigned a value of 1, ensuring focused interactions between aligned video and audio features. The attention mask is then applied directly to the attention weights produced by the self-attention mechanism, regulating modality-specific and cross-modal feature interactions by controlling how information flows between tokens. Consequently, the attention mask is a selective mechanism that suppresses irrelevant or noisy associations. Empirical analysis in table 5.5 shows that removing this mask results in a noticeable drop in localisation accuracy.

5.1.2 Score-based Contrastive Learning

In DEL, a token denotes the feature of a short temporal segment extracted from the audio or visual stream, while an instance refers to a complete video sample, as in conventional contrastive learning. Unlike instance-level approaches, DEL adopts an intra-video contrastive strategy that operates within a single video by aligning temporally corresponding audio–visual tokens and treating misaligned pairs as negatives, enabling fine-grained temporal and cross-modal discrimination for dense event localisation.

Standard contrastive learning methods operate at the instance level, often relying on heuristic positive and negative sampling across videos. While effective for aligning global representations, they are limited in capturing fine-grained temporal dynamics within a single video, especially when multiple events occur concurrently.

We adopt a score-based contrastive learning objective to improve the relationships between audio-visual features within a video during training. Unlike standard contrastive learning methods, which rely on predefined sampling heuristics, our method dynamically selects contrastive pairs using a learnt score function that ranks the audio-visual features based on contextual similarity and temporal alignment. We use both inter-sample and intra-sample objectives. The inter-sample contrastive objective strengthens the relationships between corresponding audio–video pairs across different samples in a batch, promoting cross-modal coherence, whereas the intra-sample contrastive objective operates within a single modality to capture fine-grained temporal relationships within the same video, thereby enhancing feature discrimination and alignment of audio-visual representations.

Inter-Sample Contrastive Learning Given a batch of sample pairs, we employ an inter-sample contrastive loss that maximises the cosine similarity between the $[CLS_V]$ and $[CLS_A]$ tokens of matched video-audio pairs while minimising the similarity between tokens from mismatched pairs. The $[CLS_V]$ and $[CLS_A]$ tokens are global representations of the video and audio features, respectively. The inter-sample contrastive loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{inter}} = \mathbb{E}_{z\sim[CLS_V]_p, z^+\sim[CLS_A]_p, z^-\sim I_{n\neq p}[CLS_A]_n}\ell(z, z^+, z^-) \\ + \mathbb{E}_{z\sim[CLS_A]_p, z^+\sim[CLS_V]_p, z^-\sim I_{n\neq p}[CLS_V]_n}\ell(z, z^+, z^-)\end{aligned} \tag{5.7}$$

where p and n refer to a matched video-audio pair and a negative sample from a mismatched pair, respectively. $\ell(z, z^+, z^-)$ represents the standard contrastive loss [98], defined by the following equation, and τ is a learnable temperature parameter. $\mathbb{E}_z$ denotes the expectation over the selected token pairs in the score-based contrastive learning framework, indicating that the loss is computed by averaging over the informative token pairs that the model selects dynamically, rather than by considering every possible pair.

$$\ell(z, z^+, z^-) = -\log\left(\frac{\exp(z^T \cdot z^+/\tau)}{\exp(z^T \cdot z^+/\tau) + \sum_n \exp(z^T \cdot z_n^-/\tau)}\right) \tag{5.8}$$

Intra-Sample Contrastive Learning with Score-Based Pair Selection

As shown in fig. 5.3, unlike inter-sample contrastive learning, score-based intra-sample contrastive learning enforces alignment within the same video, ensuring the model learns fine-grained feature distinctions. To achieve this, we introduce two token-level prediction mechanisms early in the network: a binary score function s_t and an event category predictor c_t. Both the score function s_t and the category predictor c_t are trained end-to-end jointly with the rest of the DEL framework. We define the score function s_t as:

$$s_t = \begin{cases} 1 & \text{if } t \text{ belongs to the ground-truth event segment} \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

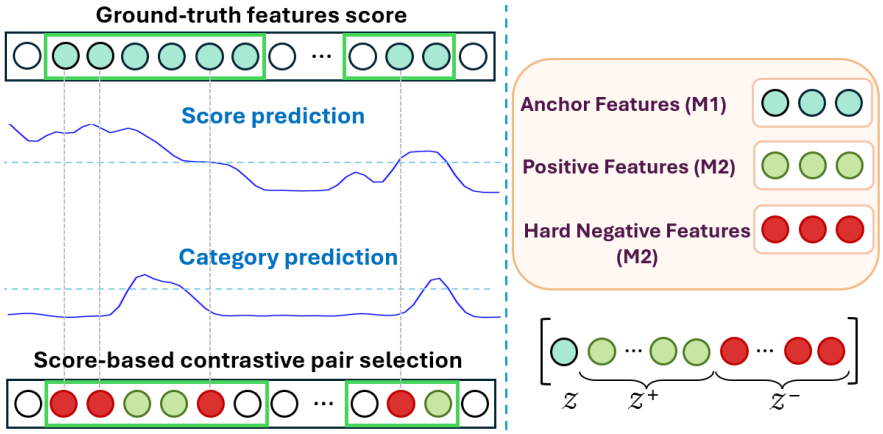


Figure 5.3: Score-based contrastive pair selection for identifying positive and hard-negative samples for each anchor within the event segment of a single modality. M1 represents the modality chosen as the anchor, while the goal is to select corresponding positive and hard-negative features from another modality represented as M2. Token-level predictions (score s_t and category c_t) provide early supervision that refines latent features. This early processing guides the contrastive selection of hard-negative samples.

This function is trained as a binary classifier using ground-truth labels derived from event annotations. DEL does not rely on explicitly annotated audio–visual token pairs; supervision is provided only through event-level temporal annotations, which are used to guide score-based pair selection. Unlike conventional hard negative mining or ranking-based approaches that operate across different videos, DEL performs score-guided hard negative selection within each video, using token-level scores to identify informative audio–visual negatives across temporal segments. Features within an annotated event segment are assigned a score of 1, while those outside are assigned zero via a binary cross-entropy loss. Our score function s_t provides early supervision to explicitly distinguish between frames that belong to annotated events and those that do not. This ensures that intra-sample contrastive pairs are meaningful, temporally aligned, and event-aware, rather than relying on random or coarse heuristics. Without explicit guidance, contrastive methods often fail to select sufficiently challenging negatives. By combining s_t with the token-level category predictor c_t, we identify hard negative features with high event confidence but incorrect category predictions. These negatives are far more informative than randomly chosen mismatched samples, forcing the model to learn subtle distinctions between visually or aurally similar events. This is particularly crucial for dense localisation, where short-duration or semantically similar events frequently co-occur.

Therefore, in parallel, the model predicts an event category c_t for each feature using a cross-entropy loss over a predefined set of event classes. By jointly training s_t and c_t at the token level, the model learns both to identify the presence of an event and to classify its action, thereby refining its latent representations before final timestamp-wise decisions.

These refined token-level predictions are key to selecting meaningful samples for the intra-contrastive loss. In DEL, contrastive pairing is restricted to a subset of high-confidence tokens identified by the score function, rather than being performed exhaustively over all possible audio–visual token pairs. Positive samples are identified as features where the model accurately predicts a score of 1 (i.e., s_t=1) and correctly classifies the event category (c_t), ensuring alignment with the ground-truth event segments across both video and audio modalities. Hard-negative samples are features that receive a high confidence score yet are assigned an incorrect event category, challenging the model to discern subtle differences.

The process results in the identification of positive visual features I_{PV}, hard-negative visual features I_{HNV}, positive audio features I_{PA}, and hard-negative audio features I_{HNA}. These samples are then utilised in an intra-sample contrastive loss function, following a similar contrastive equation defined in eq. (5.8), to enhance the model's discriminative capabilities across modalities. The intra-contrastive loss is then formulated as

$$\begin{aligned}\mathcal{L}_{\text{intra}} = \mathbb{E}_{z\sim I_{PV}, z^+\sim I_{PA}, z^-\sim I_{HNA}} \ell(z, z^+, z^-) \\ + \mathbb{E}_{z\sim I_{PA}, z^+\sim I_{PV}, z^-\sim I_{HNV}} \ell(z, z^+, z^-)\end{aligned} \tag{5.10}$$

This score-guided selection process ensures that positive samples are strongly aligned, while hard negatives provide challenging learning signals for better feature separation. Integrating hard-negative mining within intra-contrastive learning strengthens the model's ability to differentiate between visually and aurally similar events and allows representation learning and boundary prediction to reinforce each other, producing more discriminative features and more accurate temporal boundaries. This synergy makes the approach highly effective for dense event detection in untrimmed videos.

5.1.3 Path Aggregation Network for Multiscale Feature Fusion

Inspired by feature pyramid networks [301, 38, 81], we propose a path aggregation network to aggregate information across multiple temporal resolutions, preserving short-term event cues and long-term contextual dependencies. We construct a multiscale representation in which lower levels focus on fine-grained temporal details, while higher levels encode broader contextual information. Unlike the standard feature pyramid network [154], which builds static top-down feature pyramids treating hierarchies independently, we introduce two key components: (I) **Modality-Guided Adapters**, which enrich visual features with relevant audio cues and vice versa, ensuring cross-modal enrichment at each scale, rather than processing feature hierarchies independently as in standard feature pyramid networks. Our design employs both top-down and bottom-up pathways to retain lower-level, fine-grained temporal details while aggregating higher-level contextual information. These are followed by the (II) **Adaptive Pooling Module** that dynamically adjusts feature importance across different scales, emphasising crucial temporal cues, in contrast to the fixed-scale aggregation used in standard feature pyramid networks.

Modality-Guided Adapters:

As illustrated in fig. 5.4, we employ top-down and bottom-up pathways to construct a feature pyramid. This design integrates multiscale visual features $V_l \in \mathbb{R}^{T_V \times d}$, and multiscale audio features $A_l \in \mathbb{R}^{T_A \times d}$ (where $l \in \{1, 2, 3, 4, 5, 6\}$) into a series of feature pyramids (P_1 through P_n). Lower levels capture fine-grained temporal details, while higher levels encode abstract, long-term dependencies.

Adapted modality-guided adapters are strategically positioned after the top-down or bottom-up fusion processes to leverage information from both modalities for mutual enrichment and employ a max-sigmoid attention mechanism to integrate audio cues into visual features.

$$V_l' = V_l \cdot \sigma \left(\max_{u \in \{1..C_A\}} \left(V_l A_u^\top \right) \right)^\top \tag{5.11}$$

and similarly merge visual cues into audio embeddings,

$$A_l' = A_l \cdot \sigma \left(\max_{r \in \{1..C_V\}} \left(A_l V_r^\top \right) \right)^\top \tag{5.12}$$

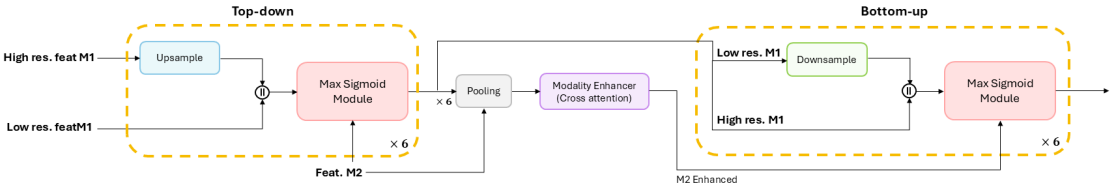


Figure 5.4: Illustration of the path aggregation network for multiscale feature fusion. The network employs a top-down and bottom-up pathway to fuse feature maps across different temporal scales, helping capture fine-grained details and high-level semantics. A modality enhancer module refines cross-modal feature representations by applying cross-attention, ensuring robust integration of audio-visual data for accurate event localisation. M1 indicates one modality (e.g., visual), and M2 represents the other (e.g., audio) .$\|$ represents the concatenation operation. The details of the max sigmoid module are shown fig. 5.5.

Here, l represents the pyramid level in the path aggregation network, while u and r index audio and visual features at that level, respectively. σ denotes the sigmoid activation function. C_A and C_V represent the number of audio and video embedding channels, respectively. The updated feature maps V_l' and A_l' are then concatenated with outputs from adjacent scales, propagating essential information throughout the pyramid.

The audio-guided adapter enriches visual features with acoustic context, while the visual-guided adapter enhances audio representations with visual semantics. These are followed by the adaptive pooling module that refines cross-modal integration by dynamically adjusting feature importance across different scales, ensuring optimal weighting of short- and long-duration event cues.

Adaptive Pooling Module (APM):

To further enhance the video features with audio information, the APM aggregates multiscale audio features into N temporal segments, resulting in audio tokens $\tilde{A} \in \mathbb{R}^{N \times d}$. We aggregate multiscale audio features into N temporal segments to establish a shared temporal resolution between audio and video modalities. Since the extracted audio and video features may have different temporal resolutions due to variations in sampling rates and processing pipelines, directly aligning them can be challenging. By fixing both modalities to N segments, we enforce

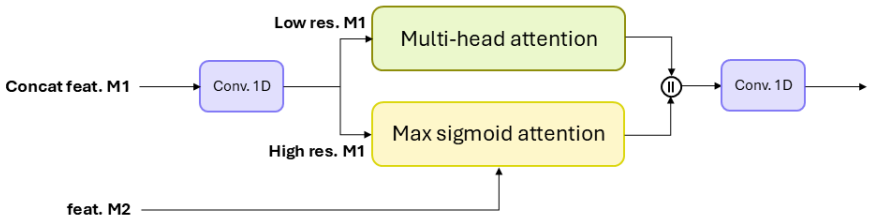


Figure 5.5: Detailed diagram of the max sigmoid module. This module is a key component of our path aggregation network, facilitating adaptive feature fusion across modalities. It takes high- and low-resolution feature maps from modality M1 and combines them with features from modality M2, using a max sigmoid function to dynamically select the most relevant features for enhanced representation learning.

a structured representation in which each token captures a comparable temporal extent across both modalities, thereby ensuring better temporal synchronisation between audio and video. This normalisation reduces modality mismatches and allows multi-head attention to operate more effectively by aligning the most relevant audio cues with corresponding video segments. Additionally, this strategy helps manage computational efficiency by reducing the number of tokens processed while maintaining rich cross-modal interactions. These tokens are then used to update the video embeddings V through multi-head attention:

$$V' = V + \text{MultiHead-Attention}(V, \tilde{A}, \tilde{A}) \tag{5.13}$$

Conversely, for audio enhancement, the APM aggregates multiscale video features into N spatial regions, producing video tokens $\tilde{V} \in \mathbb{R}^{N \times d}$. These tokens update the audio embeddings A:

$$A' = A + \text{MultiHead-Attention}(A, \tilde{V}, \tilde{V}) \tag{5.14}$$

This approach allows for efficient integration of audio context at varying temporal scales into video representations and vice versa. The resulting updated embeddings, V' and A', offer richer cross-modal representations that improve the overall performance of our audio-visual fusion network.

5.1.4 Overall Objective Function

The fused features are then passed through classification and regression heads for event category prediction and temporal boundary refinement. The final learning objective of DEL is a combination of the contrastive inter and intra losses, $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$, and the score cross entropy loss $\mathcal{L}_{score}$. Additionally, the classification head is trained using a cross-entropy loss, $\mathcal{L}_{\text{cls}}$, which ensures accurate event categorisation, while the regression head is optimised with a smooth L1 loss, $\mathcal{L}_{\text{reg}}$, to refine the temporal boundaries of each detected event:

$$\begin{aligned}\mathcal{L}_{DEL} = \lambda_1\mathcal{L}_{inter} + \lambda_2\mathcal{L}_{intra} + \lambda_3\mathcal{L}_{score} \\ + \lambda_4\mathcal{L}_{cls} + \lambda_5\mathcal{L}_{reg}\end{aligned} \tag{5.15}$$

λ values adjust the weighting between each loss term. The values are set to balance loss terms, ensuring that no single term dominates optimisation while maintaining model stability and convergence.

5.2 Datasets

To comprehensively evaluate our proposed DEL framework, we conduct experiments across four benchmark datasets that vary in complexity, modality characteristics, and temporal granularity: THUMOS14, ActivityNet 1.3, EPIC-KITCHENS-100, and UnAV-100. These datasets together provide a diverse benchmark spanning short, well-defined actions, long-duration daily activities, and complex, overlapping multimodal events in untrimmed videos. While EPIC-KITCHENS-100 was discussed in the previous chapter, it is briefly summarised here for completeness.

THUMOS14 [112] is a widely used benchmark for temporal action localisation. It includes 413 untrimmed videos from 20 action categories, with 200 validation videos used for training and 213 videos for testing. The videos primarily depict sports-related activities such as diving, pole vaulting, and basketball dunking. Each video contains multiple action instances that often overlap or occur sequentially, which makes THUMOS14 suitable for evaluating models that aim to identify precise temporal boundaries.

ActivityNet 1.3 [21] is a large-scale dataset for human activity recognition and temporal localisation. It includes more than 20,000 untrimmed videos covering 200 action categories, with 10,024 videos for training and 4,926 for validation. The dataset contains a wide variety of actions from everyday life to sports, with long-duration videos that often feature multiple activities. This makes it useful for evaluating the model's ability to handle complex temporal dynamics and multi-instance localisation.

EPIC-KITCHENS-100 [47] is a large-scale egocentric dataset that captures natural human interactions in kitchen environments. It contains approximately 100 hours of video recorded by 45 participants across 100 kitchens. The dataset provides dense temporal annotations for verbs (97 classes) and nouns (300 classes), which can be combined into composite action labels. EPIC-KITCHENS focuses on fine-grained activities such as cutting, pouring, or stirring, making it a challenging benchmark for models that must handle short and subtle actions. We use the official training and validation splits provided by the dataset.

UnAV-100 [81] is a benchmark designed for dense audio-visual event localisation. It contains 10,790 untrimmed videos with 30,059 annotated audio-visual events across 100 categories, totalling over 126 hours of video. Around 60% of the videos include multiple events, averaging 2.8 events per video, and approximately 25% contain concurrent events. The dataset spans diverse real-world scenarios, including human interactions, natural environments, and urban scenes. These characteristics make it ideal for testing a model's ability to handle overlapping and asynchronous audio-visual events.

Figure 5.6 presents examples from the four datasets to highlight their visual and structural diversity.

5.3 Experimental Details

This section provides details on implementation, including the training process and inference strategy.

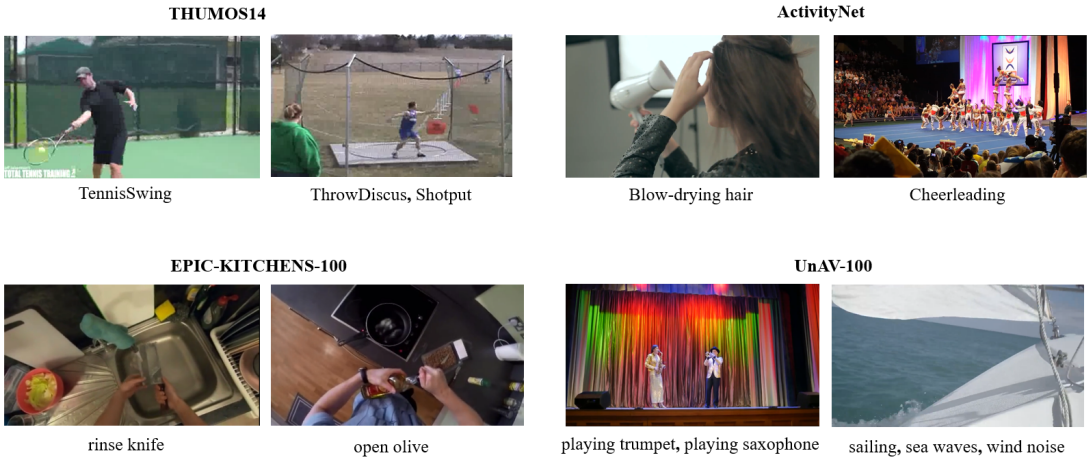


Figure 5.6: Sample frames from the four datasets used in this study.

5.3.1 Evaluation Metric: Mean Average Precision (mAP)

In temporal action localisation, mean Average Precision (mAP) is the most commonly used evaluation metric, with tIoU = 0.5 as a standard comparison reference point.

Precision (P) measures the proportion of correctly detected action instances within a single class for a given video. Specifically, for class C, precision is defined as:

$$P = \frac{TP}{TP + FP} = \frac{\text{Number of correctly predicted proposals}}{\text{Total number of predicted proposals}} \tag{5.16}$$

Since the test set contains multiple videos, Average Precision (AP) is the mean precision for a specific class C across all test videos. Further, as the test set spans multiple action classes, the mean Average Precision (mAP) is computed as the mean of AP values across all classes:

$$mAP = \frac{\sum AP}{\text{Total number of classes}} \tag{5.17}$$

In summary, under a given tIoU threshold, precision (P) quantifies the accuracy of detected action instances within a specific class in a single video, AP reflects the averaged precision across all classes in a video, and mAP generalises this across all test videos and classes. Following standard evaluation protocols, most studies report mAP at multiple tIoU thresholds to comprehensively assess model performance. In the experimental tables in this chapter, we report a column

denoted as Avg, which represents the mean Average Precision averaged over multiple temporal Intersection over Union (tIoU) thresholds.

5.3.2 Feature Extraction and Implementation Details

Our experimental setup differs across datasets due to variations in video resolution, frame rates, and feature types. Below, we provide specific details for THUMOS14, ActivityNet-1.3, EPIC-KITCHENS-100, and UnAV100. We follow existing works, and for visual feature extraction, we used I3D [25], a two-stream inflated 3D convolutional network pretrained on Kinetics-400, for THUMOS14, ActivityNet, and UnAV-100, leveraging its effectiveness in capturing spatio-temporal dynamics. Like other works [311, 217, 219] on EPIC-KITCHENS-100, where fine-grained action recognition is crucial, we employed the SlowFast network [68], pretrained on EPIC-KITCHENS. Across all experiments, audio features were extracted using the VGGish model, pretrained on AudioSet [80], thereby ensuring consistent and high-quality audio representations. Building on the general setup described above, the following outlines the dataset-specific configurations adopted in this project.

- THUMOS14: For feature extraction on THUMOS14, we utilised a two-stream I3D model [25] pretrained on Kinetics, aligning with [311, 320]. Each input clip consisted of 16 consecutive frames, processed with a sliding window stride of 4. We extracted 1024-D features from the layer preceding the final fully connected layer and combined the two-stream outputs into a 2048-D representation as input to our model. Performance was assessed using mAP across tIoU thresholds in the range [0.3:0.1:0.7], where mAP is computed by averaging the Average Precision values obtained at temporal IoU thresholds from 0.3 to 0.7 in steps of 0.1, and a prediction is considered correct at a given threshold if its temporal overlap with the ground truth exceeds that threshold. The model was trained for 50 epochs, starting with a 5-epoch linear warmup. The learning rate was initialised at 5e-4. To improve generalisation, a cosine decay schedule with a weight decay of 1e-4 was applied.

- **ActivityNet-1.3** We utilised a two-stream I3D model [25] for feature extraction, setting the sliding window stride to 16. Following previous studies [311, 153, 152], features

were downsampled to fixed lengths of 160 using linear interpolation for I3D features. Performance was measured using mAP@[0.5:0.05:0.95] and the average mAP. The model was trained for 15 epochs with a 5-epoch linear warmup, using a learning rate of 5e-4 and a weight decay of 1e-4.

- **EPIC-KITCHENS-100** We extracted features using a SlowFast network [68], which was pretrained on EPIC-KITCHENS-100 and provided by [47]. The model processed 32-frame video segments with a stride of 16, generating 2304-dimensional feature embeddings. Training was conducted on the designated training split, while evaluation was performed on the validation set.

  For evaluation, we followed the mAP@[0.1:0.1:0.5] metric and reported the average mAP, maintaining consistency with [47]. The model was trained for 30 epochs, using an initial learning rate of 5e-4 and a weight decay of 1e-4, ensuring stable convergence.

- **UnAV100** For processing UnAV-100, we extract visual and audio features while ensuring proper temporal alignment. Video frames are sampled at 25 fps, and for each segment, we use a two-stream I3D [25] network to process 24 consecutive RGB and optical flow frames. A stride of 8 is applied during feature extraction, and the outputs from both streams are concatenated to form a 2048-dimensional visual representation. Simultaneously, VGGish [102] extracts audio embeddings from 0.96-second segments with a stride of 0.32 seconds, ensuring synchronisation with the visual features. Since video lengths vary, we standardise inputs by cropping or padding sequences to a fixed length of T = 224.

  For training, we adopt the Adam optimiser and run the model for 40 epochs, incorporating a 5-epoch linear warmup at the start. The initial learning rate is set to 1e-3 and follows a cosine decay schedule for smoother optimisation. We apply a weight decay of 1e-4 to regularise training. Given that UnAV-100 focuses on temporal event localisation in untrimmed videos, we evaluate model performance using mean Average Precision (mAP). Specifically, we compute mAP at tIoU thresholds from 0.5 to 0.9 (step size 0.1) and report the average mAP over the range 0.1 to 0.9, ensuring a robust assessment of localisation accuracy.

5.3.3 Training and inferencing details

Since our framework operates on pre-extracted visual and audio features, the reported complexity and runtime correspond only to our three core modules. Together, these modules contain approximately 87M learnable parameters and require about 121 GFLOPs per 224-token input sequence. Notably, this is substantially more compact than the UnAV model, which consists of $\approx$ 139 M parameters, demonstrating that our design attains strong performance with a significantly lower computational footprint. We trained our model for 40 epochs across all datasets using NVIDIA GeForce RTX 3090, with hyperparameters set to λ_1=0.001, λ_2=1, λ_3=0.001, λ_4=1, and λ_5=1. These lambda values were set to ensure that all losses have a similar magnitude, preventing any single one from dominating the others. Runtime characteristics vary depending on sequence length and reflect dataset scale: 3.1 GPU-hours for THUMOS14, 6.1 GPU-hours for ActivityNet-1.3, 23.4 GPU-hours for EPIC-KITCHENS-100, and 9 GPU-hours for UnAV-100. Our proposed model achieves efficient inference, requiring on average $\approx$ 0.5 s per video on all datasets, with variations primarily driven by sequence length.

5.4 Results

We conduct comprehensive evaluations on four key benchmarks: THUMOS14 [112], ActivityNet-1.3 [21], EPIC-KITCHENS-100 [47], and UnAV-100 [81]. We follow the standard evaluation protocol, measuring performance using mean Average Precision (mAP) across multiple temporal Intersection over Union (tIoU) thresholds. To ensure robustness and reliability, we report averaged performance from five independent training runs to mitigate the impact of initialisation.

5.4.1 Quantitative Results

We first show the performance of our method across the range of datasets. table 5.1 shows how on THUMOS14, our method (DEL) achieves a state-of-the-art average mAP of 71.9% (mAP@[0.3:0.1:0.7]), surpassing prior methods such as TriDet by +2.6%. Notably, DEL demonstrates superior performance at higher IoU thresholds, achieving 68.4% at 0.6 and 60.5% at 0.7, significantly outperforming existing approaches. This highlights DEL's ability to localise

Method	0.3	0.4	0.5	0.6	0.7	Avg
MUSES [163]	68.9	64.0	56.9	46.3	31.0	-
ContextLoc [324]	68.3	63.8	54.3	41.8	26.2	50.9
A^2Net [295]	58.6	54.1	45.5	32.5	17.2	41.6
PBRNet [160]	58.5	54.6	51.3	41.8	29.5	-
AFSD [148]	67.3	62.4	55.5	43.7	31.1	52.0
TadTR [164]	62.4	57.4	49.2	37.8	26.3	46.6
Actionformer [311]	82.1	77.8	71.0	59.4	43.9	67.9
ASL [217]	83.1	79.0	71.7	59.7	45.8	66.8
TMaxer+MRAVFF [72]	82.2	78.2	71.5	59.9	45.3	67.4
TriDet [219]	**83.6**	**80.1**	**72.9**	62.4	47.4	69.3
DEL	81.0	78.0	71.8	**68.4**	**60.5**	**71.9**

Table 5.1: Performance comparison on THUMOS14 We report mAP across multiple tIoU thresholds and compute the average mAP. Our method outperforms previous approaches on THUMOS14 with the same feature extraction.

actions precisely. This reflects our framework's design focus on fine-grained cross-modal alignment and contrastive refinement, thereby reducing false positives and improving precision. In contrast, TriDet yields higher recall at lower IoU thresholds by producing broader temporal segments, while DEL achieves superior precision at higher thresholds through refined boundary localisation, emphasising its strength in fine-grained temporal detection.

For ActivityNet-1.3, table 5.2, DEL outperforms recent works with a +1.2% gain in average mAP, consistently improving across IoU thresholds. DEL outperforms prior methods, demonstrating its effectiveness in handling large-scale datasets with diverse and long-duration activities. The model's ability to refine event boundaries and integrate multiscale temporal cues contributes to this improvement.

On EPIC-KITCHENS-100, table 5.3, a challenging dataset focused on fine-grained cooking activities, DEL achieves 27.1% and 25.2% average mAP for the verb and noun tasks, respectively. Compared to THUMOS14 and ActivityNet, EPIC-KITCHENS-100 presents a more complex setting with lower IoU thresholds (0.1–0.5), requiring finer temporal precision and handling of subtle action variations. DEL's performance gain in this setting demonstrates its ability to distinguish short, overlapping interactions through effective cross-modal alignment. Interestingly, most approaches, including ours, perform better on verbs than nouns. Verbs form a smaller, semantically coarser label space, while nouns span thousands of fine-grained, visually similar categories that are often occluded. TIM reverses this trend, achieving stronger noun results,

Method	0.5	0.75	0.95	Avg
MUSES [163]	50.0	35.0	6.6	34.0
ContextLoc [324]	56.0	35.2	3.6	34.2
VSGN [318]	52.3	35.2	8.3	34.7
A^2Net [295]	43.6	28.7	3.7	27.8
PBRNet [160]	54.0	35.0	9.0	35.0
AFSD [148]	52.4	35.3	6.5	34.4
TadTR [164]	49.1	32.6	8.5	32.3
Actionformer [311]	53.5	36.2	8.2	35.6
ASL [217]	54.1	**67.4**	8.0	36.2
TriDet [219]	54.7	38.0	8.4	36.8
DEL	**56.9**	42.5	**14.7**	**38.0**

Table 5.2: Performance evaluation on ActivityNet 1.3. We present mAP and average mAP results across various tIoU thresholds. Our approach surpasses previous methods with the same feature extraction.

whereas DEL excels on verbs. This contrast reflects complementary inductive biases: TIM favours object-centric recognition, while DEL is optimised for dynamic actions and multimodal temporal boundaries.

Task	Method	Frozen Features	0.1	0.2	0.3	0.4	0.5	Avg
Verb	BMN [152]	SlowFast	10.8	8.8	8.4	7.1	5.6	8.4
	G-TAD [286]	SlowFast	12.1	11.0	9.4	8.1	6.5	9.4
	AF [311]	SlowFast	26.6	25.4	24.2	22.3	19.1	23.5
	ASL [217]	SlowFast	27.9	-	25.5	-	19.8	24.6
	AF + MRAV [72]	SlowFast+VGGish	27.6	26.8	25.3	23.4	19.8	24.6
	TriDet [219]	SlowFast	28.6	27.4	26.1	24.2	**20.8**	25.4
	DEL	SlowFast+VGGish	**32.2**	**29.9**	**27.8**	**25.1**	**20.8**	**27.1**
	TIM [26]	VMAE2+ASlowFast	32.9	31.6	29.6	27.0	22.2	28.6
	DEL	VMAE2+ASlowFast	**35.1**	**33.6**	**31.5**	**28.8**	**23.5**	**30.5**
Noun	BMN [152]	SlowFast	10.3	8.3	6.2	4.5	3.4	6.5
	G-TAD [286]	SlowFast	11.0	10.0	8.6	7.0	5.4	8.4
	AF [311]	SlowFast	25.2	24.1	22.7	20.5	17.0	21.9
	ASL [217]	SlowFast	26.0	-	23.4	-	17.7	22.6
	AF + MRAV [72]	SlowFast+VGGish	26.4	25.4	23.6	21.2	17.4	22.8
	TriDet [219]	SlowFast	27.4	26.3	24.6	22.2	18.3	23.8
	DEL	SlowFast+VGGish	**29.5**	**28.4**	**26.2**	**22.9**	**19.3**	**25.2**
	TIM [26]	VMAE2+ASlowFast	**36.4**	**34.8**	**32.1**	**28.7**	**22.7**	**31.0**
	DEL	VMAE2+ASlowFast	33.1	31.3	29.3	26.1	20.8	28.1

Table 5.3: Performance on the EPIC-KITCHENS-100 validation set across multiple tIoU thresholds, with average mAP reported. Our method outperforms all baselines by a significant margin using the same feature extraction, except for nouns with VMAE2+ASlowFast, where we are comparable.

Method	0.5	0.6	0.7	0.8	0.9	Avg
VSGN [318]	24.5	20.2	15.9	11.4	6.8	24.1
TadTR [164]	30.4	27.1	23.3	19.4	14.3	29.4
ActionFormer [311]	43.5	39.4	33.4	27.3	17.9	42.2
UnAV [81]	50.6	45.8	39.8	32.4	21.1	47.8
DEL	**53.4**	**48.1**	**42.6**	**35.6**	**26.9**	**51.1**

Table 5.4: Performance on the UnAV-100 test set, showcasing our method's significant improvement over all baselines using the same feature extraction. We report mAP and average mAP at various tIoU thresholds.

Finally, on UnAV-100, table 5.4, a dataset featuring complex multi-event scenarios and significant audio-visual overlap, DEL achieves a state-of-the-art average mAP@[0.1,0.1,0.9] of 51.1%, outperforming previous methods by 3.3%. The model's ability to capture cross-modal dependencies and adaptively fuse features across multiple temporal scales enables the robust localisation of overlapping and concurrent events.

As the IoU threshold increases, DEL consistently exhibits improved performance compared to other methods across the four datasets, demonstrating its ability to refine temporal boundaries and focus on more confident, well-localised events. This characteristic ensures higher precision and fewer false positives, making DEL a robust choice for real-world applications that require fine-grained audio-visual event localisation.

5.4.2 Ablation Experiments

To explore the model's performance in more detail, we ran an in-depth analysis on UnAV-100. This dataset was selected due to its complex audiovisual interactions, overlapping events, and untrimmed video format. Our study reveals that removing key modules leads to a significant performance drop, whereas multiscale fusion and feature quality enhancement improve localisation accuracy.

Component Ablation

Table 5.5 presents the results of removing key components, namely the adaptive attention mechanism, score-based contrastive learning, and path aggregation module. The analysis highlights the relative performance gains achieved by each component, demonstrating the effectiveness of

AAC	SCL	PAN	0.5	0.6	0.7	0.8	0.9	Avg
×	✓	✓	51.1	45.7	41.0	34.5	25.8	49.6
✓	×	✓	51.5	44.7	38.7	33.3	25.8	49.7
✓	✓	×	51.3	45.0	39.4	33.8	25.3	49.5
✓	✓	✓	53.4	48.1	42.6	35.6	26.9	51.1

Table 5.5: Component-wise ablation study, evaluating the individual contributions of our proposed Adaptive Attention for Cross-Modal Alignment (AAC), Score-Based Contrastive Learning (SCL), and Path Aggregation Network for Multiscale Feature Fusion (PAN) modules, on the UnAV-100 dataset

DEL in improving generalisability and refining event localisation. Qualitative comparisons of these component variants are provided in fig. 5.7.

Visualising the Impact of Each Component

Following the component ablation on UnAV-100 in table 5.5, we further illustrate the role of each module through qualitative examples in fig. 5.7. To better understand why the proposed combination of modules is effective, we visualise three video examples comparing the full model against ablated versions in achieving robust, precise event localisation, particularly in challenging scenarios with dense, overlapping audio-visual events. These examples were specifically chosen to illustrate subtle and complex interactions where unimodal baselines (audio-only, video-only) and ablated variants of DEL (removing adaptive attention, contrastive loss and path aggregation network) predict boundaries that diverge from the ground truth in distinctive ways. By analysing the predictions on temporally fine-grained classes like ”horse clip-clop,” ”raining,” ”thunder,” and ”driving motorcycle,” the complementary and synergistic benefits of fusing audio and visual modalities become evident.

In the first case, the driving motorcycle example demonstrates a long-duration, continuous event with both audio (engine noise) and visual cues (motion cues, background changes). The unimodal models either miss fine temporal details. Video-only underestimates the duration, while audio-only overextends or confuses background noise with the target activity. Without path aggregation or contrastive supervision, the predictions fluctuate, while the absence of the adaptive alignment leads to temporal drift beyond the event window. Only the complete model aligns with the annotated ground truth, underlining the necessity of combining adaptive alignment, multiscale fusion, and contrastive objectives.

In the second example, horse clip-clop, the audio-only and video-only models capture portions of the event but drift at the boundaries, while removing alignment or contrastive learning from the full model further worsens temporal precision. In contrast, the full model maintains consistent onset and offset predictions, showing that DEL enables synchronisation across modalities and compensates for modality-specific ambiguities (e.g., hoof sounds overlapping with background visual motion), while contrastive learning pushes the model to sharpen distinctions between similarly structured events. The absence of the path aggregation backbone produces the largest error, with a clearly overestimated duration, confirming that multiscale temporal fusion is essential for capturing both fine onsets and long-range dependencies.

The final example, which contains overlapping rain and thunder events, highlights the complementary roles of audio and visual cues: thunder segments are better captured by audio, whereas rain relies more heavily on visual evidence. Without adaptive attention or the path aggregation network, thunder intervals are fragmented or excessively extended, indicating that cross-modal alignment and hierarchical multiscale context are necessary for separating co-occurring events. Contrastive learning further sharpens boundaries by preventing misclassification of acoustically or visually similar segments across different times.

Collectively, these examples show how each component contributes to reducing drift, preventing confusion between overlapping events, and leveraging the complementary strengths of audio and vision for reliable dense localisation. The adaptive alignment module synchronises modalities for coherent cross-stream reasoning, contrastive learning enforces temporal discrimination across overlapping or similar events, and the path aggregation backbone provides multiscale temporal context to stabilise long- and short-duration predictions. Together, they form a complementary system that significantly outperforms individual components.

Pyramid Levels

The path aggregation module fuses features across multiple scales within our DEL framework. Table 5.6 details the results of evaluating our model on the UnAV-100 dataset using varying levels for visual and audio. Our analysis reveals that utilising six pyramid levels for both modalities yields the best performance. These results suggest a critical balance; too few levels limit the capture of multiscale context, while excessive levels introduce redundant or noisy information.

Figure 5.7: Qualitative results illustrating the effect of each component. We show ground-truth (GT) and event categories alongside predictions from audio-only (A), visual-only (V), the full audio-visual model (*AV, combining all three modules: AAT, SCL, and PAN), and ablated variants of the full AV model –AAT (without Adaptive Attention for Cross-Modal Alignment), –SCL (without Score-Based Contrastive Learning), and –PAN (without the Path Aggregation Network for Multiscale Feature Fusion). The complete AV model achieves the most precise boundaries by leveraging complementary cues and the synergy of all three modules.

L	0.5	0.6	0.7	0.8	0.9	Avg
1	47.5	42.7	37.3	30.6	22.0	45.5
2	48.5	43.2	37.7	31.2	23.0	46.4
4	48.4	43.5	38.2	32.0	23.5	47.2
6	53.4	48.1	42.6	35.6	26.9	51.1
7	51.0	45.9	40.1	33.9	24.6	49.0

Table 5.6: Ablation study on the design of the feature pyramid. L shows the number of layers for both audio and video.

Table 5.6 reinforces that multiscale audio-visual fusion and optimisation of pyramid level design are critical for robust action localisation performance.

Adapting Pyramid Depth to Dataset Characteristics

To analyse the effect of pyramid design across datasets, we extend our ablation study to THUMOS14. Unlike UnAV-100, THUMOS14 contains longer untrimmed videos, resulting in a higher maximum sequence length, which is the number of temporal feature tokens extracted from the longest videos after uniform sampling. As shown in table 5.7, we observe that the optimal number of pyramid levels depends on the dataset's temporal characteristics. While six pyramid levels yield the best performance on UnAV-100 (see Table 6 in the main paper), THUMOS14 achieves its highest accuracy at seven levels. This suggests that longer sequences benefit from a deeper temporal hierarchy, which provides richer multiscale temporal context. Using fewer levels limits the ability to capture long-range dependencies, while excessive levels may introduce redundancy or noise. It is important to note that all results reported in the main paper were obtained using six levels for consistency across datasets. Nonetheless, this analysis highlights that DEL can be tuned to dataset-specific properties to further boost performance, demonstrating its adaptability rather than fragility.

L	0.3	0.4	0.5	0.6	0.7	Avg
5	78.2	73.1	71.8	66.1	58.4	68.3
6	81.00	78.0	71.8	68.4	60.5	71.9
7	82.1	78.9	73.1	69.5	60.9	72.5
8	80.3	77.5	70.0	66.9	58.9	70.9

Table 5.7: Ablation study on the number of pyramid levels L for THUMOS14. In contrast to UnAV-100, where six levels perform best (table 5.6), THUMOS14 achieves the highest performance with seven levels, owing to its longer sequence length.

Cross-Modal Design vs. Late Fusion

To validate DEL's design, we implemented a late-fusion baseline. DEL outperforms this baseline on UnAV-100 (table 5.8), confirming that the gains stem from our alignment and aggregation mechanisms, not merely from audio access.

Method	0.5	0.6	0.7	0.8	0.9	Avg
late-fusion baseline	44.3	39.7	35.4	29.3	22.1	42.4
DEL	**53.4**	**48.1**	**42.6**	**35.6**	**26.9**	**51.1**

Table 5.8: DEL vs. late-fusion baseline on UnAV-100.

Enhancing Feature Quality

To achieve precise event localisation, the model must learn high-quality feature representations that capture modality-specific and cross-modal information. Enhancing feature quality ensures better differentiation between similar events and increases robustness against temporal inconsistencies, ultimately leading to more accurate and reliable predictions. Therefore, to explore the influence of the audio-visual feature extractors, we replace I3D with DINOv2 [189] for video and MERT-v1 [145] for audio. The results shown in table 5.9 demonstrate that integrating DINOv2 for video and MERT for audio improves performance slightly, making our proposed method suitable for use with multiple feature extractor models.

Features	0.3	0.4	0.5	0.6	0.7	Avg
THUMOS14						
I3D+Vggish	81.0	78.0	71.8	68.4	60.5	71.9
DINOv2+MERT	81.5	79.1	73.1	70.8	64.6	73.3
UnAV-100						
I3D+Vggish	53.4	48.1	42.6	35.6	26.9	51.1
DINOv2+MERT	55.0	49.7	44.1	37.4	28.4	52.7

Table 5.9: Evaluation on THUMOS14 and UnAV-100 incorporating DINOv2 for video features and MERTv1 for audio features.

Impact of Different Input Modalities A core contribution of this work is the effective fusion of audio and visual modalities within the DEL framework, enabling robust event localisation in complex, untrimmed videos. As shown in table 5.10, DEL achieves a significantly superior average mAP between 1 and 2% across the datasets when leveraging both audio and video inputs.

This represents a substantial performance increase compared to configurations using only video, demonstrating the critical and complementary role of audio-visual information for accurate event localisation. This performance gain is consistently observed across a comprehensive range of IoU thresholds. These results validate the design choices in DEL and highlight the importance of multiscale cross-modal perception and dependency modelling for advancing audio-visual scene understanding.

Data	V	A	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.75	0.95	Ave
THUMOS14	✓	×	-	-	79.9	76.8	70.7	67.6	59.7	-	-	70.8
	✓	✓	-	-	**81.0**	**78.0**	**71.8**	**68.4**	**60.5**	-	-	**71.9**
ActivityNet 1.3	✓	×	-	-	-	-	53.7	-	-	40.1	13.8	35.8
	✓	✓	-	-	-	-	**56.9**	-	-	**42.5**	**14.7**	**38.0**
EPK100 (Verb)	✓	×	30.9	28.7	26.6	24.1	20.1	-	-	-	-	26.0
	✓	✓	**32.2**	**29.9**	**27.8**	**25.1**	**20.8**	-	-	-	-	**27.1**
EPK100 (Noun)	✓	×	28.0	27.0	24.9	21.6	18.5	-	-	-	-	23.9
	✓	✓	**29.5**	**28.4**	**26.2**	**22.9**	**19.3**	-	-	-	-	**25.2**

Table 5.10: DEL performance with various modality combinations. Fusing audio and video yields the best results, emphasising the importance of multimodal input.

5.4.3 Qualitative Results.

In fig. 5.8, we present qualitative results demonstrating the effectiveness of our DEL framework in comparison to unimodal baselines. For example (a), our model, leveraging audio and visual streams, accurately localises events such as wind noise, cars passing by, driving a motorcycle, and skidding, even in scenarios with overlapping or short-duration occurrences. In contrast, the audio-only model struggles with visually-driven events like skidding, while the vision-only model fails to detect sound-based events like wind noise. Moreover, relying solely on audio leads to incorrect predictions, such as misclassifying the scene as engine knocking due to the absence of visual context. Similarly, the vision-only model, lacking critical audio cues, misinterprets the scene as auto racing from start to finish solely on visual cues. This shows the impact of audio in disambiguating visually similar activities. Distinguishing between ”man speaking” and another potential sound source in the third scenario (c) is only possible with audio input, as visual information alone is insufficient. In (b), a crowded scene, “people cheering” is missed by the vision-only model, but it is correctly identified when combined with the audio.

Similarly, in example (d), the scene is crowded, making it challenging to infer ”people cheering” solely from visual cues. The vision-only model struggles to recognise this category, while the audio modality provides crucial information for its correct identification. Additionally, detecting ”people slapping” relies primarily on visual cues. These results highlight how integrating audio and visual streams leads to more accurate and robust event localisation, particularly in complex multimodal scenarios. These results highlight the complementary nature of audio and visual modalities for precise dense event localisation, particularly in complex, real-world scenarios where events are often overlapping and context-dependent.

Figure 5.8: Qualitative results of our DEL framework for audio-visual event localisation. We present ground truth (GT) and event categories alongside predictions from audio-only (A), visual-only (V), and audio-visual (AV) models. The AV model (DEL) achieves more accurate event localisation by effectively leveraging both modalities, while the unimodal models struggle with events that rely on cross-modal cues.

5.5 Conclusion

The Dense Event Localisation framework introduced in this paper represents a significant advancement in multimodal audio-visual understanding, addressing the challenges of dense semantic action localisation in untrimmed videos. By leveraging adaptive attention mechanisms and multiscale feature fusion, DEL successfully aligns audio and visual features while preserving fine-grained temporal structures. This approach ensures robust event detection and classification, even in scenarios with overlapping events or asynchronous modalities. The integration of score-based intra- and inter-sample contrastive learning further enhances feature discrimination and cross-modal coherence, enabling the model to effectively distinguish between similar events occurring at different times.

The proposed framework demonstrates state-of-the-art performance across multiple benchmarks, including UnAV-100, THUMOS14, ActivityNet 1.3, and EPIC-KITCHENS-100. Notable improvements in average mAP scores highlight DEL's efficacy in dense event localisation tasks. By addressing limitations in existing methods such as modality misalignment and insufficient temporal modelling, DEL provides a comprehensive solution that bridges the gap between audio-visual representation learning and real-world applications. Its ability to dynamically fuse audio and visual information at multiple temporal scales ensures precise localisation of both short-duration and long-duration events.

In conclusion, DEL's innovative design paves the way for future research in multimodal video understanding, offering a scalable and efficient framework for dense event localisation. The incorporation of adaptive attention mechanisms, path aggregation networks, and dynamic contrastive learning strategies sets a new benchmark for multimodal interaction modelling. As video data continues to grow in complexity, DEL's contributions will be instrumental in advancing applications such as video retrieval, surveillance, and human activity recognition, ensuring accurate and comprehensive analysis of real-world scenes.

Chapter 6

Conclusions and Future Work

This chapter concludes the thesis by summarising the key findings, insights, and contributions from the three core works presented throughout this research. The overarching goal of the thesis was to advance the field of video understanding by developing efficient, interpretable, and multimodal learning frameworks capable of capturing both the dynamic and semantic richness of real-world visual data.

Through a progressive exploration of three key stages encompassing motion-aware self-supervised representation learning, language-grounded semantic modelling, and multimodal event localisation, this research demonstrates that integrating motion, language, and audio cues yields a more comprehensive, semantically grounded understanding of video content. Each framework addressed a distinct yet interconnected challenge: learning from unlabelled data, infusing semantics without manual supervision, and capturing temporally dense and fine-grained audio–visual interactions across untrimmed videos.

The following sections revisit these contributions, outlining each framework's main findings and identifying key limitations and directions for future research.

6.1 Conclusions

This thesis set out to achieve the following key objectives:

- Develop motion-aware self-supervised learning methods that explicitly focus on video's most dynamic and informative regions, enabling more effective representation learning without manual annotation.
- Develop a self-supervised framework that aligns visual and linguistic representations in a shared semantic space, allowing the model to learn conceptually meaningful and interpretable video features that capture meaningful visual–language associations and support robust downstream learning.
- Design a multimodal framework for dense event localisation that integrates audio and visual cues, effectively modelling complex cross-modal interactions and multi-scale temporal hierarchies to achieve fine-grained temporal reasoning and robust performance on untrimmed, real-world videos.

Together, these objectives guided the research presented in the following chapters.

In Chapter 3, we introduced MOFO, a motion-focused self-supervised framework that highlights the fundamental role of motion in video representation learning. The framework addressed a key limitation of existing self-supervised methods by guiding the model to concentrate on a scene's most dynamic and informative regions rather than relying on random masking or static visual cues. To achieve this, MOFO employed an unsupervised motion area detection method based on optical flow derivatives and motion boundaries, allowing it to identify meaningful motion while reducing the influence of camera movement and background noise. These motion-sensitive regions informed a masking strategy that encouraged the model to learn from action-relevant areas during pretraining. Beyond pretraining, MOFO extended its focus on motion into the finetuning stage via a multi-head cross-attention mechanism that fused embeddings from within and outside the detected motion area, further enhancing temporal reasoning and contextual understanding. Together, these components enabled MOFO to learn motion-aware, semantically rich representations, thereby improving accuracy and interpretability in video understanding. This work demonstrated that explicitly encoding motion provides a strong foundation for more effective and interpretable self-supervised learning in dynamic visual environments.

Chapter 4 introduced FILS, a self-supervised framework designed to enrich video representations by incorporating semantic guidance from natural language. Building on the motion-aware foun-

dation established in the previous chapter, FILS shifted the learning objective from mere pixel reconstruction to feature prediction within a semantic language space. The framework leveraged automatically generated video captions and introduced ActCLIP, a patch-wise video–text contrastive learning strategy focusing on action-relevant regions detected using motion maps from MOFO. By constructing a semantic language space via aligning visual and textual representations, FILS enabled the model to bridge motion dynamics with high-level conceptual understanding. This synergy between motion-aware learning and linguistic guidance yielded semantically grounded, interpretable representations that improved downstream performance while remaining computationally efficient. FILS demonstrated that language can be a powerful form of self-supervision, guiding visual models to develop a higher-level, transferable understanding of video data and marking an important step toward semantic video representation learning without manual annotation.

Finally, in Chapter 5, we addressed the real-world challenge of densely localising multiple, concurrent audio–visual events in long, untrimmed videos. The proposed DEL framework was designed to model complex temporal structures and cross-modal dependencies, enabling precise localisation of overlapping and asynchronous events in dynamic video scenes. DEL introduced a unified multimodal architecture that effectively integrates visual and auditory information for fine-grained temporal reasoning. The framework employs an adaptive cross-modal attention mechanism to dynamically align audio and visual streams, ensuring temporal coherence and robustness to modality asynchrony and background noise. A score-based dual contrastive learning strategy automatically identifies informative positive and hard-negative pairs to enhance discriminative capability, improving event discrimination within and across samples. Furthermore, a multi-scale path aggregation network fuses hierarchical temporal features, enabling the model to capture local motion details and high-level contextual dependencies. Together, these components enable DEL to precisely detect and classify overlapping or asynchronous audio–visual events, achieving state-of-the-art performance across multiple benchmark datasets using the same feature encoders while maintaining a compact and efficient design.

6.2 Future Research Directions

Building upon the methodologies developed in MOFO, FILS, and DEL, several compelling avenues for future research emerge, aiming to further bridge existing gaps and move towards truly comprehensive and human-like video analysis.

Generalisation and Adaptation to Unseen Scenarios

The current frameworks show strong performance on established benchmarks. However, to truly robustify these systems, future work should focus on extending their capabilities to handle unseen categories through few-shot or advanced self-supervised learning techniques. Leveraging the rich semantic space learnt by FILS could enable inference about novel actions or events from limited examples, or using MOFO's motion-focused pretraining might allow better adaptation to new environments and activities without extensive pretraining. Developing models capable of zero-shot or few-shot event detection and action recognition based on learnt semantic associations is crucial for more adaptable AI.

From Motion Saliency to Contextual Motion Understanding

While MOFO successfully introduced motion maps to focus self-supervised learning on meaningful movement and reduce the influence of camera motion and noise, its understanding of motion remains primarily low-level. Future research could explore semantic motion modelling, in which motion maps are informed by contextual or learnt cues, such as object interactions, scene dynamics, or intent. Incorporating semantic awareness into motion-guided self-supervision would enable models to distinguish between relevant human actions and irrelevant background movement, paving the way toward more intelligent and context-aware video representation learning.

End-to-End Trainable Feature Extractors

Like many previous approaches, the DEL framework in this thesis relies on pre-extracted features, which can cause feature misalignment due to inherent domain shifts and differing pretraining objectives of the feature encoders. A critical future direction is to develop fully end-to-end trainable feature extractors within these frameworks, allowing the entire model pipeline to be jointly optimised from raw pixel and audio data, potentially yielding more tightly integrated and discriminative representations and further reducing modality mismatches.

Integration of Language-Based Cues for Multimodal Event Localisation

While DEL excels in audio-visual fusion, it does not explicitly incorporate language-based cues. Conversely, FILS has successfully demonstrated the power of leveraging language semantics for enriched video representation learning. A promising direction would be to integrate FILS-like semantic language space learning directly into the DEL framework for multimodal event localisation, which could provide significantly richer contextual understanding. This is particularly valuable for disambiguating complex multi-event scenarios where audio-visual cues alone might be insufficient.

Beyond Localisation: Towards Causal and Relational Understanding

In Chapter 5, we focused on identifying what events occur and when they happen. The next frontier is to understand why these events occur and how they relate to each other. Future research could explore causal and relational reasoning frameworks, such as graph-based temporal models, causal transformers, or neuro-symbolic representations, to model dependencies between events. For instance, 'picking up a knife' is a prerequisite for 'chopping a carrot', which requires modelling complex interactions between actors, objects, and their temporal context over long temporal windows.

In summary, this thesis has explored how integrating motion awareness, semantic grounding, and multimodal reasoning can bring us closer to human-like video understanding and enable richer and more comprehensive video representations. Building upon low-level motion analysis in MOFO, progressing through language-grounded feature prediction in FILS, and culminating in dense audio–visual event localisation in DEL, this research demonstrates the value of structured learning strategies that bridge visual appearance, temporal dynamics, and cross-modal interactions. Collectively, these contributions lay the foundation for future systems capable of deeper perceptual understanding, contextual reasoning, and more intelligent interpretation of the complex visual world.

Bibliography

[1] Arif Akar, Ufuk Umut Senturk, and Nazli Ikizler-Cinbis. MAC: mask-augmentation for motion-aware video representation learning. In *BMVC*, 2022.

[2] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021.

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017.

[5] Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18794–18803, 2025.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[7] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023.

[8] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[9] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[10] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[11] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla. Hear me out: Fusional approaches for audio augmented temporal action localization. *arXiv preprint arXiv:2106.14118*, 2021.

[12] Federico Baldassarre and Hossein Azizpour. Towards self-supervised learning of global and object-centric representations. *arXiv preprint arXiv:2203.05997*, 2022.

[13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[14] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.

[15] Arnab Barua, Mobyen Uddin Ahmed, and Shahina Begum. A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions. *Ieee access*, 11:14804–14831, 2023.

[16] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[17] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[18] Shuai Bi, Zhengping Hu, Mengyao Zhao, Hehao Zhang, Jirui Di, and Zhe Sun. Continuous frame motion sensitive self-supervised collaborative network for video representation learning. *Advanced Engineering Informatics*, 56:101941, 2023.

[19] William Brendel and Sinisa Todorovic. Learning spatiotemporal graphs of human activities. In *2011 International Conference on Computer Vision*, pages 778–785. IEEE, 2011.

[20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[21] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[23] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[24] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

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

[26] Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. Tim: A time interval machine for audio-visual action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18153–18163, 2024.

[27] Shuning Chang, Pichao Wang, Fan Wang, Hao Li, and Zheng Shou. Augmented transformer with adaptive graph for temporal action proposal generation. In *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis*, pages 41–50, 2022.

[28] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139, 2018.

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

[30] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357, 2017.

[31] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[33] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020.

[34] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *ECCV*, 2022.

[35] Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15095–15104, 2023.

[36] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022.

[37] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with a long-memory transformer. In *European Conference on Computer Vision*, pages 503–521. Springer, 2022.

[38] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024.

[39] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[40] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020.

[41] Subhabrata Choudhury, Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Guess what moves: Unsupervised video and image segmentation by anticipating motion. *BMVC*, 2022.

[42] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[43] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[44] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[45] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[46] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020.

[47] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[48] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[49] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 833–842, 2019.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, volume 1, pages 4171–4186. Association for Computational Linguistics, 2019.

[53] Anxhelo Diko, Danilo Avola, Bardh Prenkaj, Federico Fontana, and Luigi Cinque. Semantically guided representation learning for action anticipation. In *European Conference on Computer Vision*, pages 448–466. Springer, 2024.

[54] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9726, 2022.

[55] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.

[56] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[57] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.

[58] Keval Doshi, Amanmeet Garg, Burak Uzkent, Xiaolong Wang, and Mohamed Omar. A multimodal benchmark and improved architecture for zero shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2021–2030, 2024.

[59] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[60] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4013–4022, 2021.

[61] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 768–784. Springer, 2016.

[62] Victor Escorcia, Ricardo Guerrero, Xiatian Zhu, and Brais Martinez. Sos! self-supervised learning over sets of handled objects in egocentric action recognition. In *ECCV*, 2022.

[63] David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector Santos-Villalobos, Rohith MV, and Xinyu Li. Motion-guided masking for spatiotemporal representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5619–5629, 2023.

[64] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021.

[65] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. *Proceedings of the Scandinavian Conference on Image Analysis*, pages 363–370, 2003.

[66] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[67] Christoph Feichtenhofer, haoqi fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022.

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

[69] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017.

[70] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[71] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, 2017.

[72] Edward Fish, Jon Weinbren, and Andrew Gilbert. Multi-resolution audio-visual feature fusion for temporal action localization. *arXiv preprint arXiv:2310.03456*, 2023.

[73] David J. Fleet and Allan D. Jepson. Stability of phase information. *IEEE TPAMI*, 1993.

[74] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.

[75] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020.

[76] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[77] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12155–12163, 2024.

[78] Kirill Gavrilyuk, Mihir Jain, Ilia Karmanov, and Cees GM Snoek. Motion-augmented self-training for video recognition at smaller scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10429–10438, 2021.

[79] Davi Geiger and Alan Yuille. A common framework for image segmentation. *International Journal of Computer Vision*, 6(3):227–243, 1991.

[80] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[81] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22942–22951, 2023.

[82] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022.

[83] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10406–10417, 2023.

[84] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022.

[85] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.

[86] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253, 2007.

[87] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[88] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d:

Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[89] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[90] Hanyu Guo, Wanchuan Yu, Suzhou Que, Kaiwen Du, Yan Yan, and Hanzi Wang. Video-to-task learning via motion-guided attention for few-shot action recognition. *arXiv preprint arXiv:2411.11335*, 2024.

[91] Sheng Guo, Zihua Xiong, Yujie Zhong, Limin Wang, Xiaobo Guo, Bing Han, and Weilin Huang. Cross-architecture self-supervised video representation learning. In *CVPR*, 2022.

[92] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022.

[93] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[94] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *NeurIPS*, 2020.

[95] Masashi Hatano, Ryo Hachiuma, and Hideo Saito. Emag: ego-motion aware and generalizable 2d hand forecasting from egocentric videos. In *European Conference on Computer Vision*, pages 119–136. Springer, 2024.

[96] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14867–14878, 2023.

[97] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[98] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[99] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[100] Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. Clip-s4: Language-guided self-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11207–11216, 2023.

[101] Rajat Hebbar, Digbalay Bose, Krishna Somandepalli, Veena Vijai, and Shrikanth Narayanan. A dataset for audio-visual sound event detection in movies. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[102] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[103] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[104] Wenxuan Hou, Guangyao Li, Yapeng Tian, and Di Hu. Toward long form audio-visual video understanding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–26, 2024.

[105] Jian Hu, Dimitrios Korkinof, Shaogang Gong, and Mariano Beguerisse-Diaz. Vismap: Unsupervised hour-long video summarisation by meta-prompting. *arXiv preprint arXiv:2504.15921*, 2025.

[106] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022.

[107] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13493–13504, 2023.

[108] Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision*, pages 724–740. Springer, 2022.

[109] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7199–7208, 2021.

[110] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[111] Minyoung Hwang, Joey Hejna, Dorsa Sadigh, and Yonatan Bisk. Motif: Motion instruction fine-tuning. *IEEE Robotics and Automation Letters*, 2025.

[112] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[113] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[114] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[115] Xun Jiang, Xing Xu, Zhiguo Chen, Jingran Zhang, Jingkuan Song, Fumin Shen, Huimin Lu, and Heng Tao Shen. Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 719–727, 2022.

[116] Licheng Jiao, Yuhan Wang, Xu Liu, Lingling Li, Fang Liu, Wenping Ma, Yuwei Guo, Puhua Chen, Shuyuan Yang, and Biao Hou. Causal inference meets deep learning: A comprehensive survey. *Research*, 7:0467, 2024.

[117] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[118] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[119] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021.

[120] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5492–5501, 2019.

[121] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications*, 2020.

[122] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 2022.

[123] Jin-Young Kim, Soonwoo Kwon, Hyojun Go, Yunsung Lee, Seungtaek Choi, and Hyun-Gyoon Kim. Scorecl: augmentation-adaptive contrastive learning via score-matching function. *Machine Learning*, 114(1):12, 2025.

[124] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[125] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8940–8949, 2020.

[126] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[127] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *International conference on computer vision*, 2011.

[128] Kevin B Kwan-Loo, José C Ortíz-Bayliss, Santiago E Conant-Pablos, Hugo Terashima-Marín, and P Rad. Detection of violent behavior using neural networks and pose estimation. *IEEE Access*, 10:86339–86352, 2022.

[129] Yann Le Cun and Françoise Fogelman-Soulié. Modèles connexionnistes de l'apprentissage. *Intellectica*, 2(1):114–143, 1987.

[130] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *International conference on learning representations*, 2020.

[131] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

[132] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7331–7341, 2021.

[133] Conglong Li, Zhewei Yao, Xiaoxia Wu, Minjia Zhang, and Yuxiong He. Deepspeed data efficiency: Improving deep learning model quality and training efficiency via efficient data sampling and routing. *arXiv preprint arXiv:2212.03597*, 2022.

[134] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *NeurIPS*, 2022.

[135] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, 2019.

[136] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[137] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[138] Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Motion-focused contrastive learning of video representations. In *ICCV*, 2021.

[139] Xiang Li, Heqian Qiu, Lanxiao Wang, Hanwen Zhang, Chenghao Qi, Linfeng Han, Huiyu Xiong, and Hongliang Li. Challenges and trends in egocentric vision: A survey. *arXiv preprint arXiv:2503.15275*, 2025.

[140] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

[141] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.

[142] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022.

[143] Yiming Li, Zhifang Guo, Xiangdong Wang, and Hong Liu. Advancing multi-grained alignment for contrastive language-audio pre-training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7356–7365, 2024.

[144] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018.

[145] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*, 2023.

[146] Han Liang, Jincai Chen, Fazlullah Khan, Gautam Srivastava, and Jiangfeng Zeng. Audio-visual event localization using multi-task hybrid attention networks for smart healthcare systems. *ACM Transactions on Internet Technology*, 2024.

[147] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

[148] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3320–3329, 2021.

[149] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019.

[150] Jiayi Lin, Shitong Sun, and Shaogang Gong. Gridclip: One-stage object detection by grid-level clip representation learning. *Pattern Recognition*, page 112187, 2025.

[151] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[152] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019.

[153] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[154] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[155] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. *arXiv preprint arXiv:2303.08914*, 2023.

[156] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006. IEEE, 2019.

[157] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 34:11449–11461, 2021.

[158] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[159] Hong Liu, Guanghui Wang, and Zhenhua Hu. Cross-modal video retrieval: A benchmark and baseline. *Information Sciences*, 460:292–304, 2018.

[160] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11612–11619, 2020.

[161] Shuo Liu, Weize Quan, Chaoqun Wang, Yuan Liu, Bin Liu, and Dong-Ming Yan. Dense modality interaction network for audio-visual event localization. *IEEE Transactions on Multimedia*, 25:2734–2748, 2022.

[162] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11915–11925, 2021.

[163] Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip HS Torr. Multi-shot temporal event localization: a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12596–12606, 2021.

[164] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022.

[165] Xiaoming Liu and Tsuhan Cheng. Video-based face recognition using adaptive hidden markov models. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[166] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, June 2022.

[167] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[168] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 344–353, 2019.

[169] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[170] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[171] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[172] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 81, pages 674–679, 1981.

[173] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. In *WACV*, 2020.

[174] Dezhao Luo, Shaogang Gong, Jiabo Huang, Hailin Jin, and Yang Liu. Generative video diffusion for unseen novel semantic video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5847–5855, 2025.

[175] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23045–23055, 2023.

[176] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[177] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

[178] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016.

[179] Gabriel Maldonado, Armin Danesh Pazho, Ghazal Alinezhad Noghre, Vinit Katariya, and Hamed Tabkhi. Moclip motion-aware fine-tuning and distillation of clip for human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2931–2941, 2025.

[180] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017.

[181] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[182] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022.

[183] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Proposal-free temporal action detection via global segmentation mask learning. In *European Conference on Computer Vision*, pages 645–662. Springer, 2022.

[184] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

[185] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022.

[186] Jingcheng Ni, Nan Zhou, Jie Qin, Qian Wu, Junqi Liu, Boxun Li, and Di Huang. Motion sensitive contrastive learning for self-supervised video representation. In *European Conference on Computer Vision*, pages 457–474. Springer, 2022.

[187] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. Egocentric vision-based action recognition: A survey. *Neurocomputing*, 2022.

[188] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[189] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[190] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.

[191] Zexu Pan, Gordon Wichern, François G Germain, Aswin Subramanian, and Jonathan Le Roux. Late audio-visual fusion for in-the-wild speaker diarization. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 174–178. IEEE, 2024.

[192] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021.

[193] AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In *European Conference on Computer Vision*, pages 76–94. Springer, 2022.

[194] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[195] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.

[196] Arjun Prashanth, SL Jayalakshmi, and R Vedhapriyavadhana. A review of deep learning techniques in audio event recognition (aer) applications. *Multimedia Tools and Applications*, 83(3):8129–8143, 2024.

[197] Rui Qian, Yeqing Li, Zheng Xu, Ming-Hsuan Yang, Serge Belongie, and Yin Cui. Multimodal open-vocabulary video classification via pre-trained vision and language models. *arXiv preprint arXiv:2207.07646*, 2022.

[198] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021.

[199] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

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

[201] Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. *Computer Vision and Image Understanding*, 235:103764, 2023.

[202] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021.

[203] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4372–4376. IEEE, 2020.

[204] Merey Ramazanova, Victor Escorcia, Fabian Caba, Chen Zhao, and Bernard Ghanem. Owl (observe, watch, listen): Audiovisual temporal context for localizing actions in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4879–4889, 2023.

[205] Kanchana Ranasinghe and Michael S Ryoo. Language-based action concept spaces improve video self-supervised learning. *Advances in Neural Information Processing Systems*, 36:74980–74994, 2023.

[206] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023.

[207] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[208] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on international conference on machine learning*, pages 833–840, 2011.

[209] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6740–6750, 2024.

[210] Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, Heysem Kaya, and Alexey Karpov. Audio-visual compound expression recognition method based on late modality fusion and rule-based decision. *arXiv preprint arXiv:2403.12687*, 2024.

[211] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, 2005.

[212] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[213] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[214] A H. Shabani, J S. Zelek, and D A. Clausi. Robust local video event detection for action recognition. In *NeurIPS, Machine Learning for Assistive Technology Workshop*, 2010.

[215] Amir Hossein Shabani, David A Clausi, and John S Zelek. Improved spatio-temporal salient feature detection for action recognition. In *BMVC*, 2011.

[216] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020.

[217] Jiayi Shao, Xiaohan Wang, Ruijie Quan, Junjun Zheng, Jiang Yang, and Yi Yang. Action sensitivity learning for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13457–13469, 2023.

[218] Dingfeng Shi, Qiong Cao, Yujie Zhong, Shan An, Jian Cheng, Haogang Zhu, and Dacheng Tao. Temporal action localization with enhanced instant discriminability. *arXiv preprint arXiv:2309.05590*, 2023.

[219] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023.

[220] Hedvig Sidenbladh. Detecting human motion with support vector machines. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 188–191. IEEE, 2004.

[221] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018.

[222] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[223] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 2014.

[224] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[225] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[226] S Sowmyayani and P Arockia Jansi Rani. Stharnet: Spatio-temporal human action recognition network in content based video retrieval. *Multimedia Tools and Applications*, 2022.

[227] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019.

[228] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[229] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

[230] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6420–6429, 2023.

[231] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, and Chuang Gan. Masked motion encoding for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2235–2245, 2023.

[232] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 1985.

[233] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[234] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021.

[235] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13526–13535, 2021.

[236] Kevin Tang, Li Fei-Fei, and Daphne Koller. Learning latent temporal structure for complex event detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1250–1257. IEEE, 2012.

[237] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018.

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

[239] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[240] Elahe Vahdani and Yingli Tian. Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4302–4320, 2022.

[241] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *TPAMI*, 2017.

[242] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[243] Namrata Vaswani, Amit K Roy-Chowdhury, and Rama Chellappa. " shape activity": a continuous-state hmm for moving/deforming shapes with application to abnormal activity detection. *IEEE Transactions on Image Processing*, 2005.

[244] Satvik Venkatesh, David Moffat, and Eduardo Reck Miranda. You only hear once: a yolo-like algorithm for audio segmentation and sound event detection. *Applied Sciences*, 12(7):3293, 2022.

[245] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289, 2018.

[246] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[247] Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 2005.

[248] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, 2019.

[249] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *ECCV*, 2020.

[250] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11804–11813, 2021.

[251] Lei Wang and Piotr Koniusz. Self-supervising action recognition by statistical moment and subspace descriptors. In *ACMMM*, 2021.

[252] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[253] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *CVPR*, 2021.

[254] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[255] Lining Wang, Haosen Yang, Wenhao Wu, Hongxun Yao, and Hujie Huang. Temporal action proposal generation with transformers. *arXiv preprint arXiv:2105.12043*, 2021.

[256] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

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

[258] Qiang Wang, Yanhao Zhang, Yun Zheng, and Pan Pan. Rcl: Recurrent continuous localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13566–13575, 2022.

[259] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14733–14743, 2022.

[260] Shiguang* Wang, Zhizhong* Li, Yue Zhao, Yuanjun Xiong, Limin Wang, and Dahua Lin. denseflow. `https://github.com/open-mmlab/denseflow`, 2020.

[261] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[262] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *AAAI*, 2020.

[263] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021.

[264] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018.

[265] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[266] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[267] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[268] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022.

[269] Yuetian Weng, Zizheng Pan, Mingfei Han, Xiaojun Chang, and Bohan Zhuang. An efficient spatio-temporal pyramid transformer for action detection. In *European Conference on Computer Vision*, pages 358–375. Springer, 2022.

[270] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, 2019.

[271] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.

[272] Shikui Wu, Hau San Wong, and Zhiwen Yu. Multi-stream deep networks for person to person violence detection in videos. *Pattern Recognition*, 57:233–247, 2016.

[273] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019.

[274] Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. Astra: An action spotting transformer for soccer videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 93–102, 2023.

[275] Kun Xia, Le Wang, Sanping Zhou, Gang Hua, and Wei Tang. Dual relation network for temporal action localization. *Pattern Recognition*, 129:108725, 2022.

[276] Kun Xia, Le Wang, Sanping Zhou, Nanning Zheng, and Wei Tang. Learning to refactor action and co-occurrence features for temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893, 2022.

[277] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018.

[278] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[279] Fanyi Xiao, Joseph Tighe, and Davide Modolo. Maclr: Motion-aware contrastive learning of representations for videos. In *European conference on computer vision*, pages 353–370. Springer, 2022.

[280] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020.

[281] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[282] Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid. M&m mix: A multimodal multiview transformer ensemble. *arXiv preprint arXiv:2206.09852*, 2022.

[283] Yuwen Xiong, Mengye Ren, Wenyuan Zeng, and Raquel Urtasun. Self-supervised representation learning from flow equivariance. In *ICCV*, 2021.

[284] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019.

[285] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3893–3901, 2020.

[286] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10156–10165, 2020.

[287] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 279–286, 2020.

[288] Cheng Xue, Xionghu Zhong, Minjie Cai, Hao Chen, and Wenwu Wang. Audio-visual event localization by learning spatial and semantic co-attention. *IEEE Transactions on Multimedia*, 25:418–429, 2021.

[289] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

[290] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3333–3343, 2022.

[291] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, 2020.

[292] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021.

[293] Haosen Yang, Deng Huang, Bin Wen, Jiannan Wu, Hongxun Yao, Yi Jiang, Xiatian Zhu, and Zehuan Yuan. Self-supervised video representation learning with motion-aware masked autoencoders. *arXiv preprint arXiv:2210.04154*, 2022.

[294] Haosen Yang, Deng Huang, Bin Wen, Jiannan Wu, Hongxun Yao, Yi Jiang, Xiatian Zhu, and Zehuan Yuan. Motionmae: Self-supervised video representation learning with motion-aware masked auto encoders. 2024.

[295] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020.

[296] Shusheng Yang, Yixiao Ge, Kun Yi, Dian Li, Ying Shan, Xiaohu Qie, and Xinggang Wang. Rils: Masked visual reconstruction in language semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23304–23314, 2023.

[297] Xitong Yang, Xiaodong Yang, Sifei Liu, Deqing Sun, Larry Davis, and Jan Kautz. Hierarchical contrastive motion learning for video action recognition. *arXiv preprint arXiv:2007.10321*, 2020.

[298] Zhao Yang, Yansong Tang, Luca Bertinetto, Hengshuang Zhao, and Philip HS Torr. Hierarchical interaction network for video object segmentation from referring expressions. In *BMVC*, 2021.

[299] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[300] Jiashuo Yu, Ying Cheng, and Rui Feng. Mpn: Multimodal parallel network for audio-visual event localization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[301] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6241–6249, 2022.

[302] Keunwoo Peter Yu. VideoBLIP.

[303] Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, and Joyce Chai. Efficient in-context learning in vision-language models for egocentric videos. *arXiv preprint arXiv:2311.17041*, 2023.

[304] Ye Yuan, Qin-Bao Song, and Jun-Yi Shen. Automatic video classification using decision tree method. In *Proceedings. International Conference on Machine Learning and Cybernetics*, volume 3, pages 1153–1157. IEEE, 2002.

[305] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019.

[306] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[307] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *CVPR*, 2022.

[308] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, 2007.

[309] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7094–7103, 2019.

[310] Can Zhang, Yuexian Zou, Guang Chen, and Lei Gan. Pan: Persistent appearance network with an efficient motion cue for fast action recognition. In *ACMMM*, 2019.

[311] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022.

[312] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019.

[313] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In *ACMMM*, 2021.

[314] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Siow Mong Rick Goh. Parallel attention network with sequence matching for video grounding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 776–790, 2021.

[315] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020.

[316] Jiayi Zhang and Weixin Li. Multi-modal and multi-scale temporal fusion architecture search for audio-visual video parsing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3328–3336, 2023.

[317] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

[318] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021.

[319] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019.

[320] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 539–555. Springer, 2020.

[321] Yue Zhao and Philipp Krähenbühl. Training a large video model on a single machine in a day. *arXiv preprint arXiv:2309.16669*, 2023.

[322] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023.

[323] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[324] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13516–13525, 2021.