Video Understanding by Design: How Datasets Shape Architectures and Insights

Lei Wang, Piotr Koniusz, Yongsheng Gao

***Abstract*—Video understanding has advanced rapidly, fueled by increasingly complex datasets and powerful architectures. Yet existing surveys largely classify models by task or family, overlooking the structural pressures[1] through which datasets guide architectural evolution. This survey is the first to adopt a dataset-driven perspective, showing how motion complexity, temporal span, hierarchical composition, and multimodal richness impose inductive biases that models should encode. We reinterpret milestones, from two-stream and 3D CNNs to sequential, transformer, and multimodal foundation models, as concrete responses to these dataset-driven pressures. Building on this synthesis, we offer practical guidance for aligning model design with dataset invariances while balancing scalability and task demands. By unifying datasets, inductive biases, and architectures into a coherent framework, this survey provides both a comprehensive retrospective and a prescriptive roadmap for advancing general-purpose video understanding.**

***Index Terms*—Video understanding, datasets, architectures, transformers, multimodal learning, procedural reasoning, temporal modeling, spatiotemporal representation, inductive bias.**

I. Introduction

THE last two decades have witnessed video understanding evolve from a niche research frontier into a cornerstone of computer vision, powering applications in surveillance, autonomous driving, robotics, healthcare, education, and large-scale multimedia retrieval [1]–[13]. Unlike images, videos encode rich spatiotemporal dynamics, hierarchical procedures, multimodal signals, and human-object interactions, making their analysis one of the most challenging yet rewarding frontiers in artificial intelligence [14], [15]. The complexity of video understanding has driven a co-evolution of datasets, learning paradigms, and model architectures: each new dataset has presented fresh challenges [16]–[22], each paradigm has offered new strategies to learn from them [3], [14], [23]–[26], and each architectural design has embodied inductive biases suited to emerging tasks [7], [9], [27]–[32]. Figure 1 illustrates how key dataset attributes, including motion complexity, temporal span, hierarchical structure, and multimodal richness, have actively guided the evolution of video understanding from isolated action recognition to context-rich reasoning.

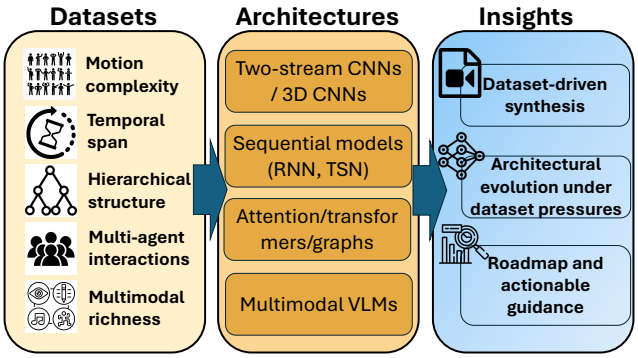


Fig. 1: Datasets as structural lenses. Key attributes, motion complexity, temporal span, hierarchical structure, and multimodal richness, impose distinct challenges that shape model design. From early motion-focused datasets fostering 3D CNNs to multimodal corpora driving transformers and vision-language models, datasets have actively guided the evolution of video understanding **from isolated action recognition to general, context-rich reasoning**.

Early progress in the field was driven by compact, manually curated datasets [33], [34] and handcrafted descriptors such as HOG3D [35] and improved dense trajectories [36]. However, the rise of large-scale datasets like Kinetics [16], Something-Something [17], ActivityNet [18], Charades [19], AVA [21], and EPIC-Kitchens [37] reshaped the landscape by introducing diverse tasks: from short-term action classification to fine-grained, long-horizon, procedural, relational, and multimodal reasoning. This dataset expansion catalyzed the transition from shallow two-stream CNNs [38] to 3D CNNs [2], temporal reasoning networks [25], transformers [27], [28], graph-based models [7], [23], [26], and multimodal vision-language frameworks [32], [39]–[43]. Parallel to dataset growth, learning paradigms have diversified beyond fully supervised training. Self-supervised learning reduced dependency on labels by using temporal consistency or masked modeling [29], [31], [44]. Few-shot and zero-shot learning enabled transfer to unseen categories [7], [45]–[51]. Reinforcement learning introduced action forecasting and policy-based video interaction [52], [53]. Most recently, VLMs and LLMs have opened the door to semantic grounding, cross-modal reasoning, and interactive video understanding, allowing systems not only to classify or

Lei Wang is with the School of Engineering and Built Environment, Electrical and Electronic Engineering, Griffith University, and also with Data61/CSIRO (email: l.wang4@griffith.edu.au).

Piotr Koniusz is with Data61/CSIRO, the School of Computing at the Australian National University (ANU), and the School of Computer Science and Engineering at the University of New South Wales (UNSW) (email: Peter.Koniusz@data61.csiro.au). Yongsheng Gao is with the School of Engineering and Built Environment, Electrical and Electronic Engineering, Griffith University (email: yongsheng.gao@griffith.edu.au).

Lei Wang and Yongsheng Gao are the corresponding authors.

[1] Structural pressures are inherent dataset constraints, *e.g.*, motion complexity, temporal span, hierarchical structure, and multimodal richness, that shape how models learn and generalize, rather than arising from class distribution, labeling schemes, or collection biases.

detect but also to answer questions, anticipate future events, and narrate procedures [14], [54], [55].

Despite extensive progress, existing surveys are fragmented. Some focus narrowly on action recognition benchmarks, others on specific model families such as CNNs, RNNs, or transformers, while more recent works highlight multimodality but lack integration of datasets, paradigms, and architectures into a unified framework [14], [15], [56]–[59]. None provide a structural lens that explains why architectures evolved as they did: namely, as responses to dataset-induced pressures and paradigm shifts. Without this integration, the field lacks a conceptual map that both contextualizes prior advances and anticipates future trends.

In this work, we present a dataset-driven survey of video understanding, emphasizing how the intrinsic structural properties of datasets, rather than chronological or superficial categorizations, drive architectural design and shape progress. By interpreting architectures as concrete responses to dataset complexity and task demands, our perspective shows a clear trajectory: from short-term clip classification to context-aware, hierarchical, relational, multimodal, and interactive reasoning. This framework not only consolidates prior work but also uncovers the underlying forces governing architectural innovation, providing actionable guidance for research and deployment. Our **key contributions** are:

i. **Filling a dataset-centric gap.** Existing surveys are largely task- or architecture-focused, overlooking how dataset characteristics and their induced inductive biases fundamentally shape model design, performance, and task suitability. We address this gap by providing a unified, dataset-driven perspective on video understanding.

ii. **Dataset-driven synthesis and analysis.** We systematically examine how dataset properties, *e.g.*, motion complexity, temporal span, hierarchical structure, and multimodal richness, guide architectural evolution. By linking these characteristics to performance trends across action recognition, temporal localization, video retrieval, and video question answering, we show recurring design patterns aligned with dataset-induced biases and the structural pressures driving model innovation.

iii. **Prescriptive roadmap and actionable guidance.** Building on these insights, we offer a task-oriented roadmap for selecting and designing models that balance temporal, relational, and multimodal reasoning with scalability and deployment constraints. We outline forward-looking directions, proposing a framework that integrates dataset properties, architectural principles, and task-specific considerations to guide the next generation of video understanding models.

By situating datasets, architectural responses, and inductive biases within a unified framework, this survey delivers both a comprehensive retrospective and a forward-looking roadmap, guiding the development of general-purpose, robust, and scalable video understanding systems.

II. Positioning Within Prior Surveys

The literature on video understanding is vast and rapidly expanding. Existing survey efforts can broadly be grouped into four complementary strands: (i) general surveys summarizing methods for action recognition and video analysis (*e.g.*, [12], [14], [15], [61], [62], [85]), (ii) focused reviews on specific architectural families (*e.g.*, CNNs, RNNs, transformers) or modalities (*e.g.*, skeleton, audio-visual, vision-language) (*e.g.*, [10], [15], [56], [64], [68], [86], [87]), (iii) evaluations and benchmark-driven comparative analyses (*e.g.*, [13], [24], [80]–[83]), and (iv) surveys on emerging paradigms such as self-supervision, generative modeling, and reinforcement/continual learning (*e.g.*, [48], [79], [88]–[92]). In this section, we group prior surveys into coherent categories and highlight how our work extends their scope by adopting a dataset-centric perspective. Our goal is not to exhaustively catalog prior work but to position representative survey efforts relative to the unique contributions of this article (see Table I).

Classical and deep-learning surveys. Early surveys focused on handcrafted features and classical pipelines (*e.g.*, spatio-temporal interest points [93], HOG3D [94], dense trajectories [95]) before tracing the transition to deep learning. Representative examples include Aggarwal and Ryoo [60] (pre-deep era) and Herath *et al*. [61], which reviewed the evolution from 2D CNNs to 3D CNNs and two-stream architectures. These works documented historical backbone evolution and early benchmark evaluations (*e.g.*, [1], [33], [34]). While foundational, these surveys were largely architecture-centric and focused on classification tasks. They overlooked how dataset structural properties (*e.g.*, temporal span, compositional complexity, multimodal richness) induce dataset-driven inductive biases in architectures and paradigms, and did not account for the influence of large-scale multimodal pretraining, VLMs, or LLMs on model evolution. In contrast, our survey systematically links dataset characteristics to both architectural and paradigm choices, showing the structural pressures that drive model innovation and task-specific performance.

Transformer and modern architecture surveys. With the rise of attention-based models, several surveys and tutorials have examined transformers in vision and video (*e.g.*, Video Transformers and ViT extension [63], [64], [96]), analyzing design patterns (*e.g.*, space-time factorization) and computational trade-offs with benchmark comparisons. While indispensable for understanding architectural design, these surveys remain narrow. They seldom consider (i) dataset-aligned inductive biases beyond attention, such as graph-based structures or generative modeling priors, (ii) interactions between architectures and learning paradigms, including how masked or multimodal objectives reshape transformers' effective biases, or (iii) dataset-driven motivations, such as which dataset properties catalyzed the shift toward transformers. Our work situates transformers within a broader taxonomy and explicitly analyzes how dataset characteristics influence their downstream utility and task-specific performance.

Multimodal and VLM/LLM surveys. As video understanding increasingly incorporates language and audio, surveys have emerged on multimodal learning (*e.g.*, video captioning, retrieval, cross-modal pretraining [14], [15], [65], [66]). These works largely catalogue datasets, alignment objectives, and evaluation protocols without explaining how dataset characteristics shape multimodal model design and paradigm se-

TABLE I: Comparison of major video understanding surveys. Unlike prior works, our survey adopts a **dataset-centric perspective**, systematically linking dataset properties to architectural design, paradigm choices, and task-driven model evolution. It provides a prescriptive roadmap for next-generation video understanding.

Representative surveys	Scope / Focus	Limitations	Our distinction / contribution
Classical & deep-learning surveys (*e.g.*, [60]–[62])	Handcrafted features; early deep pipelines (*e.g.*, 2D/3D CNNs, two-stream, *etc.*).	Architecture- and classification-centric; neglect dataset-driven design pressures.	Connect dataset properties (temporal span, motion complexity, compositionality) to architectural evolution; show how classical/deep pipelines adapt to dataset-induced inductive biases.
Transformer / modern architecture surveys (*e.g.*, [27], [63], [64])	Transformers; attention mechanisms; efficiency trade-offs.	Focus on transformers; limited discussion of dataset-induced motivations or alternative architectures.	Situate transformers in broader evolution of video models; link adoption to dataset properties and task-specific demands; highlight design patterns enabling spatiotemporal and relational reasoning.
Multimodal / VLM / LLM surveys (*e.g.*, [14], [15], [65], [66])	Video-language pretraining, cross-modal datasets, emerging video-LLMs.	Catalogue datasets/models without explaining how dataset structure drives multimodal alignment and architectural choices.	Treat multimodal alignment as dataset-driven design; analyze modality dominance and hybrid pipelines; connect VLM/LLM pretraining to structural and temporal dataset properties.
Structured representation surveys (*e.g.*, [13], [26], [67]–[72])	Skeletons, graphs, relational reasoning, multi-agent interactions.	Modality-specific; rarely link representational choices to dataset structure (*e.g.*, agent density, relational complexity).	Position graphs within unified design space; highlight how dataset properties dictate the choice of graph-based, attention-based, or convolutional representations for relational reasoning.
Self-supervised / generative / pretraining surveys (*e.g.*, [9], [73]–[79])	Contrastive, masked, generative, and hybrid pretraining objectives.	Method-centric; seldom connect pretraining choice to dataset properties and task demands.	Emphasize dataset-informed paradigm selection; show when contrastive, masked, or generative pretraining improves task-specific performance; highlight hybridization guided by temporal, compositional, and multimodal structure.
Benchmarks / dataset surveys (*e.g.*, [13], [24], [80]–[84])	Benchmark datasets, evaluation metrics, collection protocols.	Treat datasets as static; ignore structural properties and their influence on model evolution.	Introduce *datasets as structural lenses*; link temporal span, motion complexity, compositional depth, agent density, and multimodality to model design and paradigm selection.
Foundation-model surveys (*e.g.*, [10], [14])	Foundation models across video tasks and modalities.	Emphasize model scale, not dataset- or task-driven evolution.	Connect foundation models to dataset-induced pressures; integrate them into the dataset-centric taxonomy of architectures and paradigms.
Ours (2025)	Comprehensive coverage: from classical action recognition to multimodal, VLM-, and LLM-augmented video understanding.	–	First survey to unify datasets, architectures, and paradigms under a dataset- and task-centric lens; provides prescriptive guidance linking dataset structure to architectural design, paradigm adoption, and downstream performance.

lection. We advance this perspective by (i) treating multimodal alignment as a dataset-informed paradigm, reflecting invariances induced by different data regimes, (ii) analyzing modality imbalance and modality-dominance failure modes as consequences of dataset properties, and (iii) situating VLM- and LLM-augmented models within hybrid pipelines integrating generative priors and self-supervision, highlighting how dataset structure guides paradigm and architectural choices.

Structured representations. Many surveys cover skeleton-based action recognition and graph-based approaches for modeling human-object interactions, multi-agent dynamics, and relational reasoning [13], [26], [67]–[72]. While valuable, these works largely remain modality-specific and rarely consider how dataset properties influence the choice of representational paradigms. Our survey situates graph-based representations within a broader architecture and paradigm design space, showing how dataset characteristics, *e.g.*, agent density, relational complexity, compositional depth, drive the selection of graph, attention, or self-supervised representations for effective spatiotemporal and relational reasoning.

Self-supervision, generative, and pretraining surveys. Surveys on self-supervised and generative modeling [9], [73]–[79] summarize objectives, augmentations, and downstream transfer. While informative, they seldom analyze how dataset characteristics guide pretraining paradigm selection. In contrast, our survey emphasizes alignment between dataset properties and learning paradigms, highlighting when contrastive, masked, or generative objectives are most effective, and how hybridization or curriculum strategies can be guided by structural and temporal dataset properties.

Benchmarks and dataset-centered studies. Several surveys focus on benchmarks [13], [24], [80]–[84], evaluation practices, and dataset analyses (*e.g.*, Kinetics [16], EPIC-KITCHENS [22], [37]). While useful, they often treat datasets as static resources. In contrast, we adopt datasets as structural lenses, analyzing properties such as temporal span, compositionality, annotation granularity, multimodal richness, and agent density, and show how these drive paradigm and architectural choices, enabling more principled guidance for model design, pretraining, and dataset construction.

Reinforcement, continual, and privacy-aware learning. Some surveys review reinforcement learning, continual learning, and federated/privacy-preserving approaches, highlighting challenges such as catastrophic forgetting, non-i.i.d. data, and sparse rewards [97]–[102]. While valuable, these works are often disconnected from mainstream video understanding. From a dataset-centric perspective, we highlight how specific dataset properties, such as long-horizon egocentric sequences or distributed data collection, can influence the applicability of these strategies, providing guidance on when they may complement standard video understanding pipelines.

III. Datasets as Structural Drivers

Rather than treating datasets as static benchmarks, we frame them as structural lenses that actively shape what video models can learn. Motion complexity, temporal span, hierarchical

TABLE II: Datasets as structural lenses for video understanding. The table organizes major datasets by structural properties: supervision, compositionality, multi-agent density, and temporal span, and highlights the architectural advances they spurred. Row colors indicate the dataset's primary focus: motion/fine-grained actions (red), procedural/compositional tasks (green), temporal/stepwise tasks (blue), VLM/video-language tasks (purple), and mixed categories for overlaps. View: 3P = third-person, Ego = egocentric. Mods: R = RGB, F = Flow, A = Audio, D = Depth, P = Pose/Skeleton, I = IR, T = Text, M = IMU. Anno: Cls = class labels, Temp = temporal segments, ST = spatio-temporal boxes, Cap = captions, Step = procedural steps, QA = question answering, Grnd = text grounding. Struct: Amp/Span/Comp/Agents, where Amp = motion amplitude (H/M/L), Span = temporal span (S/M/L), Comp = compositionality (-/C/H = none/compositional/hierarchical), Agents = agent density (L/M/H). Impulse: 2S = two-stream, 3D = 3D CNN, TSN/TCN, TRF = transformers/attention, ST-GCN = graph/skeleton, VLM = vision-language pretraining, HOI = hand-object interaction, Det = localized detection.

Dataset	Year	Scale	View	Mods	Anno	Struct	Primary Task	Impulse
KTH actions [103]	2004	2.4K clips	3P	R	Cls	H/S/-/L	Simple actions	2S, 3D
Weizmann [104]	2005	90 clips	3P	R	Cls	H/S/-/L	Simple actions	2S
IXMAS Actions [105]	2006	1148 clips	3P (multi-view)	R	Cls	M/S/-/L	View invariance	2S
Hollywood [106]	2008	1.4K clips	3P	R	Cls	M/S/-/M	Movie actions in-the-wild	2S
Hollywood2 [107]	2009	1.7K clips	3P	R	Cls	M/S/-/M	Movie actions	2S
Collective Activity [108]	2009	44 videos	3P	R	ST (groups)	M/S/-/H	Social/group acts	GNN
Olympic Sports [109]	2010	783 clips	3P	R	Cls	H/S/-/L	Sports actions	2S
MSRAction3D [110]	2010	567 clips	Kinect frontal	D,P	Cls	M/S/-/L	Depth/skeleton actions	ST-GCN
MSRActionPairs3D [110]	2010	360 clips	Kinect frontal	R,D,P	Cls	M/S/-/L	Interaction pairs	ST-GCN
HMDB51 [34]	2011	6.8K clips	3P	R	Cls	M/S/-/M	Diverse actions	2S, 3D
UCF101 [33]	2012	13K clips	3P	R	Cls	H/S/-/L	Actions/sports	3D, TSN
UTKinect-Action3D [111]	2012	199 clips	Kinect frontal	R,D,P	Cls	M/S/-/L	Skeleton actions	ST-GCN
G3D-Gaming [112]	2012	20 classes	Kinect frontal	R,D,P	Cls	H/S/-/L	Gaming actions	ST-GCN
UCF50 [113]	2013	6.7K clips	3P	R	Cls	H/S/-/L	Actions/sports	3D
Florence3D [114]	2013	215 clips	Kinect frontal	R,D,P	Cls	M/S/-/L	Skeleton actions	ST-GCN
JHMDB [115]	2013	928 clips	3P	R,P	ST	M/S/-/M	Pose + localized acts	Det
Sports-1M [1]	2014	1M clips	3P	R	Cls	H/S/-/L	Large-scale sports	3D pretrain
Northwestern-UCLA [116]	2014	1.5K clips	Kinect multi-view	R,D,P	Cls	M/S/-/L	Cross-view actions	ST-GCN
UWA3D (I/II) [68]	2014/15	701/1.1K	Kinect multi-view	R,D,P	Cls	M/S/-/L	View-invariant depth	ST-GCN
NTU RGB+D 60 [117]	2016	56K clips	3P	R,D,P,IR	Cls	M/S/-/L	Large-scale skeleton	ST-GCN
InfAR [118]	2016	600 clips	3P	I	Cls	M/S/-/L	Infrared actions	Robustness
Thermal Simulated Fall [119]	2016	44 clips	3P	I	Cls	L/S/-/L	Fall detection (IR)	Safety
DALY [120]	2016	3.6K ann.	3P	R	ST+Temp	M/M/-/M	Daily ST actions	Det
MultiTHUMOS [121]	2016	400 vids	3P	R	Temp (dense)	M/M/C/M	Dense multilabel acts	TRF
Volleyball (group activity) [122]	2016	4830 clips	3P	R	ST (players)+Cls	M/S/-/H	Group activity	GNN, Det
NfS (object tracking) [123]	2017	100 vids	3P	R	Boxes	M/M/-/M	Object tracking	Det
Kinetics-400 [16]	2017	306K clips	3P	R	Cls	H/S/-/M	General actions	3D (I3D), TRF
AudioSet (video) [124]	2017	2M clips	3P	R,A	Weak labels	M/S/-/M	AV tagging/pretrain	AV Fusion
Kinetics-Skeleton [26]	2018	260K	3P	P	Cls	M/S/-/M	Pose-only actions	ST-GCN
AVA [21]	2018	211K ann.	3P	R,F	ST	L/M/-/H	Atomic ST actions (multi-agent)	Det, TRF
Diving48 [125]	2018	18K	3P	R,F	Cls	L/S/-/L	Fine-grained dives	TRF
Moments in Time [126]	2019	1M+	3P	R,A	Cls	M/S/-/M	Event recognition	3D, TRF
Kinetics-600/700 [16]	2018/19	496K/650K	3P	R	Cls	H/S/-/M	Scale for pretraining	3D, TRF
NTU RGB+D 120 [127]	2019	114K	3P	R,D,P,IR	Cls	M/S/-/M	Larger skeleton	ST-GCN
FineGym [128]	2020	32K	3P	R	Cls	L/S/H/L	Fine-grained hierarchy	TRF
VGGSound [129]	2020	210K clips	3P	R,A	Cls	M/S/-/M	Audio-visual events	AV Fusion
AVA-ActiveSpeaker [130]	2020	3.65M frames	3P	R,A	ST (speaker)	L/S/-/H	AV diarization	AV Fusion, Det
UAV-Human [131]	2021	22K clips	3P (UAV)	R,P	Cls	M/S/-/M	Aerial human acts	Robustness
UCF101-24 [132]	2024	24 classes	3P	R	ST	H/S/-/M	ST detection	Det, 3D
EPIC-SOUNDS [133]	2025	100h	Ego	A,R	Temp	L/M/C/M	Ego audio events	AV Fusion
Berkeley MHAD [134]	2013	660 clips	3P	R,D,P,A	Cls	M/S/-/L	Multisensor fusion	ST-GCN
Something-Something V1/V2 [17]	2017/18	108K/221K clips	3P	R	Cls	L/S/C/M	Object-centric relations	TRN, TRF
CAD-60 [135]	2011	68 clips	Kinect single-view	R,D,P	Cls	L/S/-/L	ADL+HOI (depth)	ST-GCN
GTEA Gaze [136]	2012	17 vids	Ego	R,A (gaze)	Temp	L/M/C/M	Egocentric HOI + gaze	HOI
CAD-120 [137]	2013	120 clips	Kinect frontal	R,D,P	Temp	L/M/C/M	HOI sequences (procedural)	ST-GCN, TRN
50 Salads [138]	2013	50 vids	3P	R	Temp+Step	L/M/H/L	Fine-grained cooking	RNN/TCN
Breakfast [139]	2014	77h	3P	R	Temp+Step	L/M/H/L	Procedural activities	RNN/TCN
GTEA Gaze+ [140]	2015	37 vids	Ego	R,A (gaze)	Temp	L/M/C/M	Egocentric HOI + gaze	HOI
SYSU 3D HOI [141]	2015	480 clips	3P	R,D,P	Cls	L/S/-/M	HOI (depth)	ST-GCN
EPIC-KITCHENS [37]	2018	39K clips	Ego	R,F	Temp	L/M/C/M	HOI, fine-grained egocentric	HOI, TRF
YouCook2 [142]	2018	15K segs	3P	R,T	Temp+Step	L/M/H/M	Cooking segmentation	TRF
EGTEA Gaze+ [143]	2018	28 h	Ego	R,A (gaze)	Temp	L/M/C/M	Egocentric HOI + gaze	HOI
YouCook2-BoundingBox [144]	2018	15K seg	3P	R	ST (HOI)	L/M/H/M	Obj-centric cooking	Det, HOI
COIN [145]	2019	12K vids	3P	R,T	Step	L/M/H/M	Instructional steps	TRF
CATER [146]	2019	5.5K	Synth	R	Temp	L/S/C/L	Compositional reasoning	TRF
CLEVRER [147]	2019	20K	Synth	R,T	QA	L/S/C/L	Causal reasoning	TRF
CrossTask [148]	2019	4.7K	3P	R,T	Step	L/M/H/M	Weakly sup. steps	TRF
MOMA [149]	2021	2.4K vids	3P/Ego	R,T	Hier	L/M/H/H	Multi-agent hierarchy	TRF, GNN
MOMA-LRG [150]	2022	148 h	3P/Ego	R,T	Hier	L/M/H/H	Multi-agent hierarchy	TRF, GNN
ADL [151]	2009	10h	3P	R	Cls	L/S/-/L	Household ADL	2S
MSRDailyActivity3D [152]	2012	320 clips	Kinect frontal	R,D,P	Cls	L/S/-/L	ADL (depth)	ST-GCN
MPII Cooking [153]	2012	3.7K segs	3P	R	Temp	L/M/C/M	Cooking steps (procedural)	TSN, TRN
MPII Cooking 2 [153]	2015	273 clips	3P	R	Temp	L/M/C/M	Fine-grained cooking	TRN
EPIC-KITCHENS-100 [22]	2020	90K clips	Ego	R,F,A	Temp+Step	L/M/H/M	HOI + narration	HOI, TRF, VLM
THUMOS'14 [154]	2014	24 classes	3P	R	Temp	H/M/-/L	Temporal detection	3D, TSN
ActivityNet [18]	2015	28K clips	3P	R	Temp	M/M/-/M	Temporal localization	3D, TSN
Charades [19]	2016	66K segs	3P	R,F	Temp	L/L/C/H	Overlapping indoor actions	TRF, Det
PKU-MMD I [155]	2017	1.1K clips	3P	R,D,IR,P	Temp	M/M/-/M	Multimodal detection	ST-GCN
FineAction [121]	2018	11.6K clips	3P	R	Cls	H/S/-/M	Temporal action localization	2S
Charades-Ego [20]	2018	68K	Ego+3P	R	Temp	L/M/C/H	Ego/3P alignment	TRF, HOI
SoccerNet [156]	2018	500 games	3P	R	Temp	H/M/-/M	Sports spotting	TRF
HACS [157]	2019	1.5M clips	3P	R,F	Temp	M/M/-/M	Temporal localization (large)	TRF
PKU-MMD II [158]	2020	1K	3P	R,D,IR,P	Temp	M/M/-/M	Multimodal detection	ST-GCN
BDD100K (video) [159]	2020	100K vids	3P	R,GPS,IMU	Boxes/Tracks	M/M/-/H	Driving perception/TA	Det, TRF
MSVD (YouTube2Text) [160]	2011	1.9K vids	3P	R,T	Cap	L/S/-/M	Captioning	VLM
MSR-VTT [161]	2016	200K pairs	3P	R,T	Cap	M/S/-/M	Captioning/retrieval	VLM
LSMDC (Movies) [162]	2016	128K sent.	3P	R,T	Cap	M/S/-/H	Movie-description	VLM
DiDeMo [163]	2017	10K	3P	R,T	Grnd	L/S/-/M	Moment grounding	TRF, VLM
ActivityNet Captions [164]	2017	20K vids	3P	R,T	Temp+Cap	M/M/C/M	Dense captioning	TRF, VLM
TGIF-QA [165]	2017	104K QA	3P	R,T	QA	L/S/-/M	Video QA	TRF
MSRVTT-QA [166]	2017	10K vids	3P	R,T	QA	L/S/-/M	QA	TRF
Charades-STA [167]	2017	9.8K pairs	3P	R,T	Grnd	L/M/C/H	Language grounding	TRF
TVQA [168]	2018	153K QA	3P	R,T	QA	L/M/-/H	Multimodal QA (subs)	TRF
VATEX [169]	2019	41K	3P	R,T	Cap	M/S/-/M	Multilingual captions	VLM
NExT-QA [170]	2021	5.4K vids	3P	R,T	QA	L/M/C/M	Temporal reasoning QA	TRF
AGQA [171]	2021	192M QA	3P	R,T	QA	L/M/C/M	Compositional QA	TRF
WebVid-2M/10M [172]	2021	2.5M/10M pairs	3P	R,T	Cap (weak)	M/L/-/M	VLM pretraining	VLM
HD-VILA-100M [173]	2022	100M	3P	R,T,A	Cap (weak)	M/L/-/M	Hi-res pretrain	VLM
Ego4D [174]	2022	3.7K h	Ego	R,A,T	Multi-task (NLQ, AV, gaze)	L/L/H/H	Long-term egocentric	HOI, TRF, VLM
VidChapters-7M [175]	2023	7M seg	3P	R,T	Temp+Cap	L/M/C/M	Chaptering/summary	TRF, VLM
InternVid [176]	2023	234M	3P	R,T	Cap (weak)	M/L/-/M	Video-text pretrain	VLM
Panda-70M [177]	2024	70M	3P	R,T,A	Cap (weak)	M/L/-/M	Multimodal pretrain	VLM
MiraData [178]	2024	16K h	3P	R,T,A	Cap (weak)	M/L/-/M	Multimodal pretrain	VLM
OpenVid-1M [179]	2025	1M vids	3P	R,T	Cap (weak)	M/L/-/M	VLM pretraining	VLM
OpenVidHD-0.4M [179]	2025	433K vids	3P	R,T	Cap (weak)	M/L/-/M	VLM pretraining	VLM
Koala-36M [180]	2025	36M	3P	R,T,A	Cap (weak)	M/L/-/M	Multimodal pretrain	VLM
AVA-Kinetics [181]	2020	230K ann.	3P	R	ST	M/M/-/H	Large ST localization	TRF, Det
TACOS [182]	2013	127 vids	3P	R,T	Grnd	L/M/C/M	Grounding in cooking	TRF, VLM
HowTo100M [183]	2019	136M clips	3P	R,A,T	Weak Align	L/L/H/M	Instructional pretraining	VLM, TRF

Reading guide. *Anno* shows what the dataset supervises (*e.g.*, procedural steps, spatio-temporal boxes, captions/QA). *Struct* summarizes motion, temporal, and relational aspects. *Impulse* highlights modeling advances spurred by the dataset (*e.g.*, Kinetics→I3D/Transformers; Something-Something→relation reasoning; AVA/Volleyball→ detection/graph models; EPIC/Ego4D→ HOI/egocentric transformers; HowTo100M/WebVid→ vision-language pretraining).

composition, multi-agent interactions, and multimodal signals define not only task difficulty but also the inductive biases, representational geometries, and reasoning strategies architectures should adopt. From early datasets of isolated, high-amplitude actions to procedurally rich, fine-grained, and contextually grounded corpora, each dataset introduces distinct spatiotemporal, relational, and multimodal challenges. This reframes dataset design as a driver of architectural evolution, coordinating the field's shift from narrow action recognition toward general, procedural, and multimodal video understanding, a connection rarely foregrounded in prior surveys.

A. A Dataset-Bias-Architecture Framework

We formalize this relationship as a dataset-bias-architecture framework, in which benchmark properties impose structural pressures that guide inductive biases and, ultimately, architectural design. Coarse, motion-centric datasets encouraged two-stream CNNs and 3D ConvNets optimized for local spatiotemporal cues. Procedural and long-horizon corpora demanded recurrence, temporal hierarchies, and attention-based models to capture extended workflows. Multi-agent benchmarks fostered relational reasoning through graph-based representations, while multimodal corpora drove the rise of alignment modules and large-scale video-language foundation models. Seen in this way, the historical progression of architectures can be interpreted as a sequence of responses to dataset-imposed constraints, with datasets functioning as engines that guide innovation toward greater temporal depth, relational complexity, and multimodal integration.

Table II summarizes the most popular datasets from the past 20 years, with row colors indicating their primary categories and overlaps. Early benchmarks were dominated by motion-focused datasets, while procedural, temporal, and video-language datasets have grown steadily, reflecting the community's increasing interest in fine-grained actions, stepwise reasoning, and multimodal understanding.

Below, we discuss these four main categories in detail, highlighting their evolution and the trends revealed by this collection of benchmarks.

B. Motion & Fine-Grained

Motion complexity. Motion represents the foundational signal in video understanding, providing critical cues for action discrimination, intent inference, and interaction modeling. Across datasets, the manner in which motion is captured, emphasized, and structured varies dramatically, shaping both model development and evaluation strategies. Early benchmarks, including KTH [103], Weizmann [104], HMDB51 [34], and UCF101 [33], illustrate the initial stage of dataset design, where the focus was on clearly perceivable human movements in relatively controlled environments. These datasets primarily feature high-amplitude, visually salient motions, such as running, basketball dunking, or swinging, often captured from third-person viewpoints with limited background clutter. Such design enabled early architectures, *e.g.*, 3D ConvNets, two-stream networks, and optical-flow-based models, to effectively learn coarse global motion patterns, yet provided limited challenge for modeling subtle, context-dependent, or overlapping actions [24], [38], [184], [185].

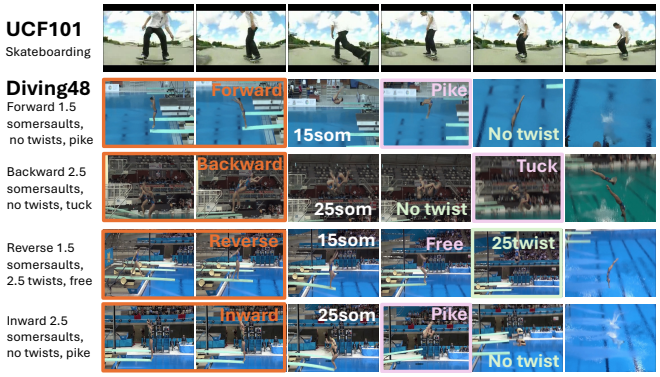


Fig. 2: Motion complexity across datasets. UCF101 (top) illustrates high-amplitude actions such as skateboarding, where global body motion dominates. Diving48 (bottom) contains four dive categories distinguished by subtle variations in rotation, twist, and posture, requiring fine-grained motion modeling. The contrast reflects the shift from early benchmarks focused on salient **whole-body movements** to modern datasets that demand recognition of nuanced **micro-motions**.

Progressing beyond coarse motion, intermediate datasets introduced more realistic variability in camera angles, scene context, and actor appearance. Hollywood [106], Hollywood2 [107], and Collective Activity [108] expose models to complex backgrounds, multi-person interactions, and social context, while datasets like MSRAction3D [110] and UTKinect-Action3D [111] provide depth and skeleton information to capture 3D body dynamics. These datasets necessitate models capable of disentangling actor motion from background variations, handling viewpoint changes, and interpreting relational motion among multiple agents. Similarly, group-activity and social interaction datasets, including Volleyball [122] and DALY [120], require tracking multiple agents simultaneously, highlighting the importance of relational reasoning and graph-structured representations for motion modeling [3], [186].

Recent datasets further emphasize fine-grained, subtle, and context-dependent motion, presenting new challenges for model design. FineGym [128] and Diving48 [125] exemplify actions where minor differences in rotational velocity, limb alignment, or entry angle define entirely distinct classes, requiring temporal precision and spatial fidelity. Here, global motion representations are insufficient; models should capture micro-motion, joint trajectories, and nuanced temporal dependencies, motivating architectures such as multi-scale 3D convolutions, temporal transformers, and pose-based representations [187]–[190]. Figure 2 illustrates this progression: UCF101 exemplifies high-amplitude, global body motions, whereas Diving48 demonstrates fine-grained distinctions in rotation, twist, and posture, highlighting the evolution from early salient-motion datasets to benchmarks that require nuanced micro-motion modeling. Similarly, AVA [21] demonstrates the

challenges of multi-person, occluded scenarios where low-amplitude gestures such as handshake or object manipulation should be disambiguated within complex visual contexts. In egocentric datasets like EPIC-KITCHENS [22], [37], motion complexity is further amplified by camera ego-motion and hand-object interactions, requiring models to decouple actor-induced motion from environmental and self-motion dynamics, a task largely absent in early third-person datasets.

Motion complexity in modern datasets is inherently multi-dimensional, capturing variations in amplitude, temporal coherence, spatial coverage, and multi-agent interactions. High-amplitude actions, such as jumping or running, contrast with low-amplitude, localized gestures like typing or stirring. Temporal coherence distinguishes cyclic patterns, such as dribbling or walking, from discrete, isolated movements. Spatial coverage differentiates whole-body motions from localized actions involving hands, facial expressions, or small objects. Multi-agent interactions introduce additional layers of complexity, requiring models to capture interdependent motions, relational dynamics, and social context. Extended temporal datasets such as Breakfast [139] and Charades [19] further demand disentanglement of overlapping motion streams, where concurrent actions like chopping vegetables while attending to boiling water should be understood as temporally and semantically distinct yet relationally dependent [191]–[193]. Collectively, these complexities underscore that motion is not a single attribute but a multi-faceted property that directly informs dataset design, feature representation, and model architecture.

Action concepts. Beyond motion itself, the semantic conceptualization of actions within datasets significantly shapes model learning, generalization, and reasoning. Early large-scale datasets such as Sports-1M [1] and Kinetics-400 [16] provide general action categories that emphasize robust recognition of prominent spatiotemporal patterns, including running, jumping, or playing instruments. These coarse labels enable models to capture broad motion trends and object-affordance cues but often fail to discriminate subtle, context-dependent differences, limiting fine-grained generalization. In contrast, specialized benchmarks such as FineGym and Diving48 demand precise discrimination of nuanced variants, requiring models to attend to detailed body alignments, rotations, and micro-temporal cues [125], [128]. First-person video datasets further integrate object semantics and environmental context, as shown in EPIC-KITCHENS, where differentiating cut tomato from cut cucumber relies not only on hand motion but also on object identity, affordances, and interaction dynamics. Atomic action datasets, including AVA, extend this challenge to multi-agent and overlapping actions, highlighting the necessity of spatiotemporal attention, multi-scale feature extraction, and graph-based relational modeling [194]–[196].

Through this lens, the evolution of motion-focused and fine-grained datasets reflects a broader trajectory in video understanding: from coarse, single-agent, high-amplitude actions to multi-agent, context-rich, and micro-motion-sensitive activities. This progression has directly shaped model design, encouraging the development of architectures capable of disentangling overlapping motion streams, reasoning over relational and temporal hierarchies, and integrating multi-modal cues. By situating model evaluation within this nuanced understanding of motion complexity and action conceptualization, researchers can better assess generalization, robustness, and zero-shot reasoning, providing insights that directly inform both dataset curation and the design of next-generation video models.

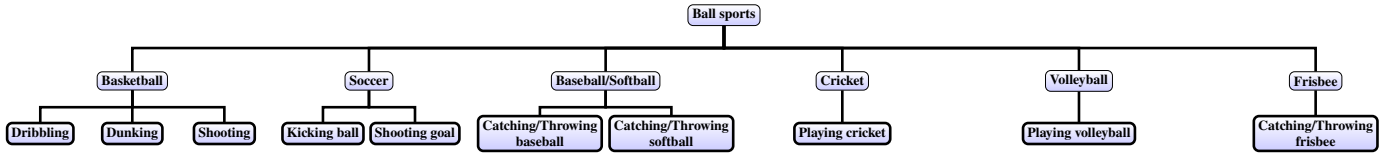
C. Procedural & Compositional

Hierarchical structure represents a fundamental dimension of video datasets that is often underexplored. Real-world actions rarely occur in isolation; they are organized both semantically and procedurally, and understanding these relationships is crucial for models that aim to generalize, perform compositional reasoning, and handle multi-step activities. Datasets differ in how they capture these hierarchies, and analyzing these differences reveals structural pressures that shape model design, evaluation, and generalization.

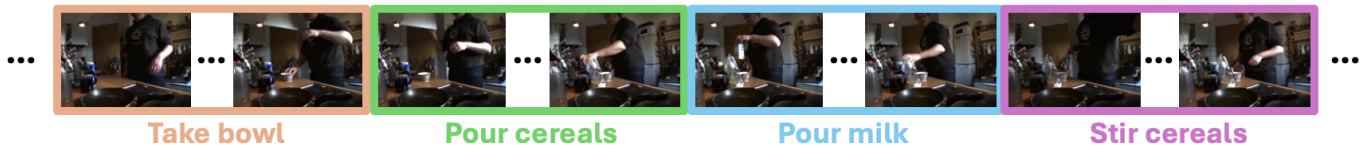
One form of hierarchy is **taxonomic**, which groups semantically related actions under broader categories. Datasets such as Sports-1M [1] and Kinetics [16] illustrate this approach: fine-grained classes like soccer, basketball, and tennis are unified under ball sports, providing shared features that models can exploit for recognition and zero-shot generalization. Hierarchically grounded taxonomies enable models to transfer knowledge across semantically similar actions, for example, recognizing handball after learning soccer and basketball, by using structural similarities in objects, motion patterns, and context [2], [24], [197]–[199]. Figure 3a visualizes a taxonomic hierarchy for ball sports in Kinetics-400, grouping related actions such as basketball dribbling, dunking, and shooting under a broader category, which supports semantic generalization and compositional reasoning.

Compositional hierarchies encode **procedural dependencies** among sub-actions that constitute complex tasks. Instructional datasets, including CAD-60 [135], GTEA Gaze [136], CAD-120 [137], 50 Salads [138], Breakfast [139], HowTo100M [183], and YouCook2 [142], capture sequences such as *open milk carton, pour milk, add cereal, stir*. By representing activities as sequences of sub-actions, these datasets encourage models to learn action primitives, reason over temporal structure, and generalize to novel procedural combinations [28], [186], [191]–[193], [200]–[216]. Larger-scale egocentric datasets such as EPIC-KITCHENS [37] extend this compositional reasoning to fine-grained hand-object interactions and continuous workflows, highlighting the need for hierarchical transformers, relational graphs, and attention mechanisms capable of capturing both short-term manipulations and long-range dependencies. Figure 3b illustrates a procedural hierarchy in Breakfast, showing how sub-actions like take bowl, pour cereals, and stir compose a complete task.

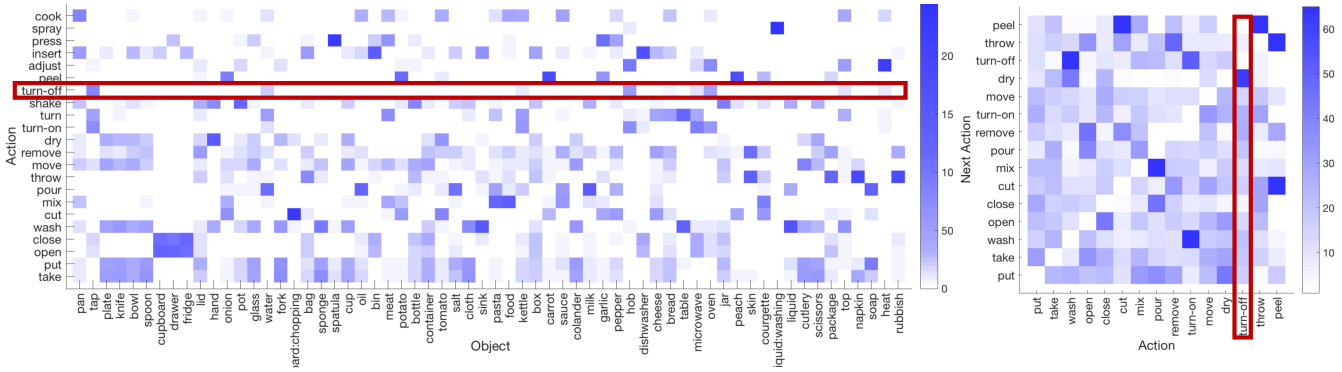
A third type of hierarchy is **contextual**, which situates actions within their environment, interacting objects, or social context. Identical motions may correspond to different semantic roles depending on context. Datasets such as AVA [21] and EPIC-KITCHENS [37] capture these distinctions: lifting a hand may mean picking up a cup, waving hello, or grabbing a dumbbell, depending on surrounding cues and task stage. Figure 3c further demonstrates contextual and



(a) Ball sport actions in Kinetics-400, grouped by category (*e.g.*, dribbling, dunking, shooting under Basketball), illustrating taxonomic hierarchies that support generalization and compositional reasoning.



(b) Making cereal in Breakfast, showing sub-actions (take bowl, pour cereals, pour milk, stir) that form the overall task.



(c) (Left) Frequently co-occurring verb-noun pairs in action segments. (Right) Next actions, excluding repetitions of the same action. Images adapted from [37]. The contextual hierarchy is evident (highlighted in red), *e.g.*, 'turn on' applied to tap, kettle, hob, microwave leads to diverse subsequent actions, *e.g.*, dry, turn on, remove, wash, *etc*.

Fig. 3: Hierarchical and compositional structures in video datasets. (a) Kinetics-400: **taxonomic hierarchy** of ball sport actions showing semantic groupings that support generalization and compositional reasoning. (b) Breakfast: **procedural activity** sequence illustrating sub-actions that compose a complete task. (c) EPIC-KITCHENS: **contextual hierarchy** of verb-object actions, highlighting how the same verb can lead to different next actions depending on context. Together, these examples emphasize the importance of hierarchical and relational structures for modeling and understanding complex human activities.

compositional hierarchies in EPIC-KITCHENS, where fine-grained hand-object interactions form continuous workflows. Modeling contextual hierarchies requires integrating visual, temporal, and environmental information, motivating multi-modal and context-aware representations [3], [29]–[31], [44], [194], [196], [200], [217]–[220].

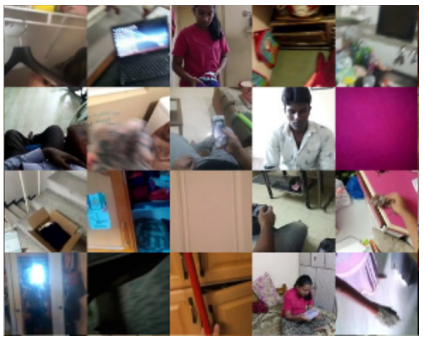
Encoding hierarchical, procedural, and contextual relationships directly influences model performance and generalization. Without such hierarchies, models may confuse visually similar sub-actions with distinct semantic roles, *e.g.*, cut carrot versus cut cucumber, or misinterpret multi-step tasks. Hierarchical annotations, whether explicit or inferred, provide the relational structure necessary to reason over atomic actions, their composition, and context. This facilitates zero-shot learning, cross-domain transfer, and the development of architectures that capture dependencies across scales, from sub-action primitives to complex, multi-agent procedures [28], [186], [191]–[193], [200]–[203], [205]–[210], [212]–[216]. Furthermore, hierarchical datasets enable graded evaluation, reflecting errors at different semantic or procedural levels, and inform the design of structured models such as graph networks, hierarchical transformers, and temporally-aware RNNs/TCNs. By systematically synthesizing procedural, compositional, and contextual hierarchies across datasets of varying scale, viewpoint, and modality, we illuminate recurring design patterns and dataset-driven pressures that have shaped modern video understanding architectures.

D. Egocentric & Long-horizon

Recent video datasets increasingly include long, continuous sequences where multiple actions unfold sequentially or concurrently, reflecting real-world activity complexity [19]–[22], [37], [139], [172], [173], [176]–[180]. For instance, Charades [19] captures everyday household activities with temporally overlapping actions, such as *pick up cup while walking to the kitchen*, requiring models to disentangle concurrent motion streams and track multiple sub-actions. Procedural datasets like Breakfast [139] provide dense temporal annotations across multi-step activities; a single breakfast video may sequentially include *cut vegetables, boil water, add pasta, stir sauce*, embedding fine-grained sub-actions within a higher-level activity. Correct interpretation demands hierarchical temporal

(a) MPII Cooking 2: third-person static view showing procedural cooking activities.

(b) EPIC-KITCHEN-100: egocentric, moving camera capturing continuous hand-object interactions, *e.g.*, open bin, get tomato, put glass, highlighting long-horizon temporal dependencies.



(c) Charades-Ego: combined third-person and egocentric views capturing overlapping everyday actions.

Fig. 4: Examples of egocentric and long-horizon video datasets highlighting temporal and procedural complexity. (a) MPII Cooking 2 shows controlled procedural cooking actions from a **third-person** perspective. (b) EPIC-KITCHENS-100 captures continuous hand-object interactions from an **egocentric** perspective, emphasizing multi-step workflows and temporal dependencies. (c) Charades-Ego combines third-person and egocentric views, showcasing overlapping everyday activities, which challenge models to reason over concurrent and sequential actions. Collectively, these datasets demonstrate how varying viewpoints, temporal granularity, and action overlap shape model design for **long-horizon** video understanding.

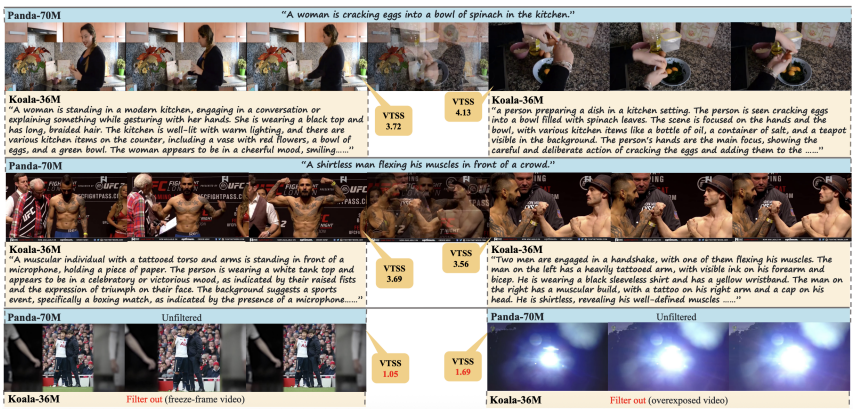
reasoning, from micro-level motions to macro-level procedural dependencies, motivating architectures with multi-scale temporal modeling [3], [25], [167], [218], [221]–[223].

Egocentric datasets further intensify temporal and relational complexity. EPIC-KITCHENS [22], [37] captures continuous hand-object interactions from a first-person perspective, where actions like *slice cucumber* or *pour milk* occur within fluid workflows including transitions, pauses, and context-dependent variations. Figure 4 shows representative examples of such datasets, including third-person procedural recordings (MPII Cooking 2 [153]), egocentric continuous interactions (EPIC-KITCHENS-100 [22], [37]), and combined third-person/egocentric views (Charades-Ego [20]), highlighting the challenges of long-horizon temporal reasoning, procedural complexity, and multi-view action overlap. Modeling such sequences requires capturing both short-term manipulations and long-range activity dependencies, highlighting the need for hierarchical transformers, recurrent attention mechanisms, or graph-based relational networks [3], [29]–[31], [44], [194], [196], [200], [217]–[220]. Ego-motion introduces additional challenges, as camera movement is tightly coupled with the actor's motion, requiring disentanglement of self-motion from object-centric interactions.

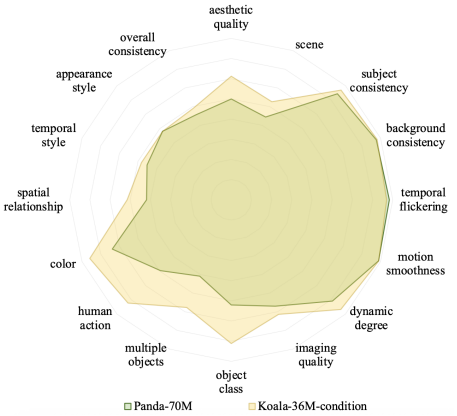
Spatial grounding is equally critical. Datasets like AVA [21] provide bounding boxes for multiple actors along with temporally localized labels, enabling models to resolve overlapping or interacting actions such as *talk to while pick up object*. Such dense spatiotemporal annotation supports multi-agent reasoning, concurrent action detection, and interaction understanding, which short-clip or flat-label datasets cannot address. Similarly, first-person and multi-view procedural datasets, *e.g.*, CAD-60 [135], GTEA Gaze [136], CAD-120 [137], 50 Salads [138], YouCook2 [142], and MOMA [149], [150], encourage architectures capable of integrating temporal hierarchies, attention to hand-object relations, and relational graph reasoning.

Temporal span, hierarchical structure, and egocentric perspective collectively introduce structural constraints that shape model design [3], [200], [217], [218]. Short-clip or single-action datasets favor frame-based or 3D convolutional networks for local motion modeling, whereas long-horizon datasets like EPIC-KITCHENS-100 [22] or YouCook2-BoundingBox [144] demand temporal transformers, memory-aware modules, and graph-based relational networks to capture long-range dependencies, sub-action composition, and multi-agent interactions [29]–[31], [44], [194], [196], [219], [220]. Procedural and instructional datasets, including Breakfast [139], COIN [145], and CATER [146], further drive hierarchical and step-aware reasoning for sequential task modeling. Egocentric datasets such as GTEA Gaze+ [140], [143] emphasize attention to gaze, hand-object relations, and perspective-specific dynamics, fostering relational modeling and context-sensitive feature integration. Temporal and spatiotemporal annotations across datasets like AVA [21], Charades-Ego [20], and PKU-MMD [155], [158] reinforce fine-grained localization capabilities, guiding architectures toward concurrent action detection and multi-agent reasoning.

Collectively, these design pressures show recurring patterns



(a) Dataset comparison: Koala-36M *vs.* Panda-70M. Koala-36M provides a larger, higher-quality dataset that improves alignment across visual, audio, and textual modalities, ensuring better temporal consistency. Beyond richer captions and more precise temporal splits, Koala-36M facilitates learning of fine-grained hand-object interactions, multi-step procedural sequences, and hierarchical action structures. Its improved filtering via the Video Training Suitability Score (VTSS) enables models to focus on high-quality, contextually informative video segments, supporting generalizable reasoning across diverse real-world activities.



(b) Quantitative evaluation of text-to-video generation on Koala-36M *vs.* Panda-70M. Models trained on Koala-36M achieve higher performance in aesthetic quality, object recognition, multi-object interactions, human action accuracy, and color consistency, showing that large-scale, multimodal datasets improve understanding of complex, overlapping, hierarchical actions.

Fig. 5: Koala-36M illustrates the power of large-scale, **multimodal** datasets in advancing video understanding compared to prior datasets such as Panda-70M. Images adopted from [180]. By combining visual, audio, and textual modalities, Koala-36M allows models to disambiguate visually similar actions, track overlapping and multi-step procedural sequences, and generalize across diverse contexts. The dataset's rich temporal structure and **detailed annotations** support hierarchical and compositional modeling, temporal reasoning, and cross-modal alignment, enabling models to capture complex human-object interactions, anticipate future actions, and perform higher-level procedural understanding. Such multimodal resources **bridge the gap between specialized action recognition and general-purpose video processing**, fostering robust, context-aware, and scalable video understanding systems.

linking dataset properties to model evolution, demonstrating how structural biases inherent in datasets actively guide innovation in video understanding architectures.

E. Multimodal Corpora

Modern video datasets increasingly recognize that visual information alone is often insufficient to fully capture the semantics of actions, particularly in complex, fine-grained, or procedural scenarios. Multimodal signals, including audio, text, and metadata, introduce complementary cues that not only improve performance but also impose structural pressures that guide model design. Audio tracks convey critical information about actions that may be visually subtle or partially occluded. For example, the sound of chopping, pouring, or clinking utensils allows models to disambiguate visually similar gestures, such as cut tomato versus cut cucumber, where visual cues alone are ambiguous. Environmental sounds, such as footsteps, machinery, or applause, provide context for temporal alignment and action recognition in real-world scenarios [124], [129], [133], [172]–[174], [176]–[180].

Textual modalities enrich video understanding by providing semantic grounding that complements visual and audio information. Instructional and narrated datasets, such as HowTo100M [183], include natural language descriptions aligned with video sequences, spanning detailed procedural steps to high-level activity summaries. These annotations enable models to map observed actions to semantic concepts, bridging low-level motion cues and high-level activity reasoning. For instance, a sequence labeled *whisk eggs and add to pan* allows a model to disambiguate visually similar sub-actions such as stir ingredients versus mix batter. Temporal reasoning is also facilitated, as textual cues often describe action sequences that extend beyond the duration of individual visual clips [144], [208], [224], [225].

Contextual information further shapes understanding by situating actions within their environment. The same motion may correspond to different actions depending on surrounding objects or participants, for example, lifting a hand may indicate pick up cup in a kitchen, wave hello in a social scene, or grab dumbbell in a gym. AVA [21] captures such nuances via spatial and contextual annotations in crowded movie scenes, whereas EPIC-KITCHENS [22], [37] records first-person hand-object interactions where object identity and placement are central to action semantics. Multi-agent scenarios emphasize relational reasoning, requiring models to track concurrent actions and interactions over space and time [224], [226], [227].

Across datasets, these multimodal properties systematically shape architectural evolution [41], [228], [229]. Small-scale captioning datasets, *e.g.*, MSVD [160] and MSR-VTT [161], encourage cross-modal embedding learning. Temporally grounded corpora, including DiDeMo [163] and ActivityNet Captions [164], enforce fine-grained alignment between visual frames and textual cues, which drives the adoption of frame-level attention and alignment mechanisms. Procedural and instructional datasets, *e.g.*, HowTo100M [183], TACOS [182], VidChapters-7M [175], and compositional QA datasets like AGQA [171], promote hierarchical and memory-aware architectures capable of capturing long-range dependencies and compositional action structure [10], [12], [14], [15], [64]–[66], [79]. Egocentric and first-person datasets, including EPIC-KITCHENS and Ego4D [174], emphasize relational modeling and attention-based reasoning for hand-object and agent-object interactions. Large-scale pretraining datasets, *e.g.*, WebVid-2M/10M [172], HD-VILA-100M [173], InternVid [176], Panda-70M [177], MiraData [178], OpenVid [179], and Koala-36M [180], enable models to generalize across tasks and support zero-shot reasoning in retrieval, captioning, temporal grounding, and question answering [133], [172], [173], [176]–[180]. Figure 5 shows the advantages of Koala-36M over prior datasets like Panda-70M, showing how large-scale, multimodal corpora with rich temporal structure and detailed annotations facilitate hierarchical action modeling, multi-step procedural understanding, and cross-modal alignment. Multimodal QA datasets, including TVQA [168], further highlight the importance of integrating dialogue, subtitles, and temporal reasoning for structured comprehension of complex video content.

Multimodal corpora show that integrating visual, audio, and textual signals systematically shapes model design, promoting hierarchical, attention-driven, and memory-aware architectures. This unified analysis across datasets highlights how multimodal richness, temporal structure, and scale collectively drive architectural innovations, fulfilling our goal of a dataset-driven synthesis of video understanding challenges.

IV. Benchmark Insights from Key Video Models

Tables III and IV benchmark representative video models across recognition, detection, retrieval, localization, and question answering. These results show how dataset properties interact with architectural choices and pretraining strategies. We organize the discussion into two parts: (i) recognition and detection benchmarks focusing on spatiotemporal representation, and (ii) multimodal tasks such as retrieval and question answering that require video-language alignment.

A. Spatiotemporal Modeling for Recognition and Detection

Short-clip, motion-centric datasets strongly favor architectures that explicitly model local spatiotemporal dynamics (see Table III). On HMDB51 and UCF101, early Two-Stream variants and 3D CNNs consistently outperform others, highlighting the importance of capturing instantaneous motion cues. For instance, Two-Stream'16 achieves 93.5% on UCF101 and 69.2% on HMDB51, while RGB-I3D reaches 95.6% and 74.8%, respectively. These results suggest that for datasets with short, well-constrained clips, exploiting frame-level motion information through two-stream fusion or early 3D convolutions remains highly effective. Long-range, compositional datasets benefit from architectures that model temporal dependencies across extended sequences. On Something-Something V1/V2 (SSv1/SSv2), sequential approaches such as TSM, TRN, and TSN show substantial gains over short-term models. For example, TSM attains 66.0% on SSv2 and 52.6% on SSv1, while TRN scores 55.5% and 42.0% on the same datasets, respectively. This emphasizes that capturing object interactions, temporal ordering, and higher-level sequence composition is critical, as simple motion cues are insufficient for datasets requiring understanding of temporal context.

Procedural and multi-agent datasets demand attention-based or transformer architectures capable of relational reasoning and large-scale context modeling. On AVA v2.2, transformer and self-supervised models such as MaskFeat and VideoMAE achieve 38.8-42.6 mAP, whereas traditional 3D CNNs underperform or are not reported. Epic-Kitchens-100 further highlights this trend: TSM achieves 38.3% on Action, 67.9% on Verb, and 49.0% on Noun, while Motionformer reaches 44.5%, 67.0%, and 58.5%, respectively. These results indicate that procedural or relational tasks benefit from architectures that integrate long-range temporal dependencies with feature attention and pretraining on large datasets.

Large-scale, generic action recognition datasets demonstrate that performance scales with model capacity and pretraining sophistication. On Kinetics-400/600/700, modern transformers (Swin, MViT, VideoMAE, InternVideo) consistently outperform early 3D CNNs, achieving top-1 accuracies above 80% and up to 84% (InternVideo and InternVideo2 on K700). This trend shows that dataset scale, model capacity, and self-supervised pretraining collectively influence performance, emphasizing the need for careful model selection when moving from small, constrained datasets to large, diverse video collections. Cross-dataset patterns show a clear alignment between dataset characteristics and architectural design. Early 3D CNNs and two-stream networks excel at short, motion-sensitive clips, sequential models dominate compositional and interaction-heavy datasets, and transformer/self-supervised models achieve the best results on procedural, multi-agent, or densely annotated benchmarks. This alignment provides a practical roadmap for model selection: matching architectural inductive biases to the temporal, relational, and compositional properties of the target dataset maximizes performance while guiding design choices for scalability.

B. Multimodal Alignment for Retrieval and Reasoning

Early backbone architectures such as I3D, R(2+1)D, and SlowFast consistently excel in temporal action localization, achieving strong mAP on THUMOS'14 (up to 66.8%) and moderate performance on ActivityNet and HACS (see Table IV). These results highlight their ability to capture fine-grained spatiotemporal motion patterns. However, their lack of reported performance on retrieval or question answering benchmarks (*e.g.*, MSR-VTT, MSVD-QA) highlights their limited capacity for multimodal reasoning or language grounding. This establishes a baseline: temporal modeling alone is

TABLE III: Top-1 performance of representative video models on **action recognition and detection** datasets. For each dataset, the best-performing model variant is reported. Datasets: HMDB51 and UCF101, action classification (top-1 accuracy); Sports-1M, with Clip Hit@1 (C@1), Video Hit@1 (V@1), and Video Hit@5 (V@5) metrics; AVA v2.1 and v2.2, action detection (mAP); Diving48, Moments in Time (Moments), Kinetics-400/600/700 (K400/K600/K700), Something-Something v1/v2 (SSv1/SSv2), and ActivityNet (ANet), top-1 classification accuracy; Charades, action detection (mAP); Epic-Kitchens-100, top-1 accuracy reported separately for Action, Verb, and Noun. Model abbreviations: 2S indicates Two-Stream, F-ST-ConvNet indicates Factorized Spatio-Temporal ConvNet, and NL indicates Non-Local. A dash (-) indicates results not reported for a dataset. Early two-stream and 3D CNNs excel on short clips, sequential models (TSN, TRN, TSM, SlowFast) capture long-range and compositional actions, and transformer-based models dominate procedural, multi-agent, and relational tasks.

Model	**HMDB51**	**UCF101**	**Sports-1M**			**AVA**$_{v2.1}$	**AVA**$_{v2.2}$	**Diving48**	**Moments**	**K400**	**K600**	**K700**	**SSv1**	**SSv2**	**EK100**			**ANet**	**Charades**
			C@1	V@1	V@5										Action	Verb	Noun		
Slow Fusion [1]	-	65.4	41.9	60.9	80.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Two-stream'14 [38]	59.4	88.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.9	-
Two-stream'16 [184]	69.2	93.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F-ST-ConvNet [230]	59.1	88.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Conv Pooling [197]	-	88.6	70.8	72.4	90.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C3D [2]	-	90.4	46.1	61.1	85.2	-	-	-	-	-	-	-	-	-	-	-	-	65.8	10.9
RGB-I3D [24]	74.8	95.6	-	-	-	14.5	-	-	29.5	71.1	71.9	-	45.8	-	-	-	-	-	35.5
Flow-I3D [24]	77.3	96.7	-	-	-	-	-	-	-	63.4	-	-	-	-	-	-	-	-	-
2S I3D [24]	80.9	98.0	-	-	-	15.6	-	-	-	74.2	75.7	-	-	-	-	-	-	-	-
P3D ResNet [198]	-	93.7	47.9	66.4	87.4	-	-	-	-	-	-	-	-	-	-	-	-	75.1	-
R(2+1)D RGB [199]	74.5	96.8	57.0	73.0	91.5	-	-	-	-	74.3	-	-	-	-	-	-	-	-	-
R(2+1)D Flow [199]	76.4	95.5	46.4	68.4	88.7	-	-	-	-	68.5	-	-	-	-	-	-	-	-	-
2S R(2+1)D [199]	78.7	97.3	-	73.3	91.9	-	-	-	-	75.4	-	-	-	-	-	-	-	-	-
S3D [231]	75.9	96.8	-	-	-	-	-	-	-	77.2	-	-	48.2	-	-	-	-	-	-
X3D [218]	-	-	-	-	-	-	27.4	-	-	79.1	81.9	-	-	-	-	-	-	-	47.1
NL RGB-I3D [25]	-	-	-	-	-	-	-	-	-	77.7	-	-	-	-	-	-	-	-	37.5
TSN [185]	71.0	94.9	-	-	-	-	-	-	50.1	-	-	-	30.0	33.2	60.2	46.0	89.6	-	-
TRN [221]	-	83.8	-	-	-	-	-	-	28.3	-	-	-	42.0	55.5	35.3	65.9	45.4	-	25.2
TSM [232]	73.6	95.9	-	-	-	-	-	-	-	74.3	-	-	52.6	66.0	38.3	67.9	49.0	-	-
SlowFast [3]	-	-	-	-	-	27.3	30.7	77.6	-	79.8	81.8	71.0	-	63.1	38.5	65.6	50.0	-	45.2
TVN [233]	75.5	-	-	-	-	-	-	-	30.7	-	-	-	-	-	-	-	-	-	54.6
MoViNet [234]	-	-	-	-	-	-	-	-	40.2	81.5	84.8	72.3	-	64.1	47.7	72.2	57.3	-	63.2
ECO [235]	72.4	94.8	-	-	-	-	-	-	-	70.0	-	-	49.5	-	-	-	-	-	-
VTN [236]	-	-	-	-	-	-	-	-	37.4	79.8	-	-	-	-	-	-	-	-	-
AssembleNet [237]	-	-	-	-	-	-	-	-	34.3	-	-	-	-	-	-	-	-	-	58.6
TimeSformer [28]	-	-	-	-	-	-	-	81.0	-	80.7	82.2	-	-	62.5	-	-	-	-	-
ViViT [27]	-	-	-	-	-	-	-	-	38.5	84.9	85.8	-	-	65.9	44.0	66.4	56.8	-	-
MViT [223]	-	-	-	-	-	-	28.7	-	-	81.2	83.8	-	-	68.7	-	-	-	-	47.7
Motionformer [238]	-	-	-	-	-	-	-	-	-	81.1	82.7	-	-	68.1	44.5	67.0	58.5	-	-
Swin [239]	-	-	-	-	-	-	-	-	-	84.9	86.1	-	-	69.6	-	-	-	-	-
MaskFeat [44]	-	-	-	-	-	37.8	38.8	-	-	87.0	88.3	80.4	-	75.0	-	-	-	-	-
MViTv2 [240]	-	-	-	-	-	-	33.5	-	-	86.1	87.9	79.4	-	73.3	-	-	-	-	-
VideoMAE [29]	73.3	96.1	-	-	-	-	39.3	-	-	87.4	-	-	-	75.4	-	-	-	-	-
VideoMAE V2 [31]	88.1	99.6	-	-	-	-	42.6	-	-	90.0	89.9	-	68.7	77.0	-	-	-	-	-
InternVideo [30]	89.3	-	-	-	-	-	-	-	-	91.1	91.3	84.0	70.0	77.2	-	-	-	94.3	-
InternVideo2 [32]	80.7	97.3	-	-	-	-	-	-	51.2	92.1	91.9	85.9	-	77.5	-	-	-	95.9	-

insufficient for tasks that require semantic alignment across video and text. The emergence of video-language pretrained models marks the next major shift. Architectures such as CLIP4Clip, Frozen, VIOLET, and ALPRO achieve substantial gains on retrieval benchmarks, with MSR-VTT R@1 reaching 32.5–42.1%. These gains show that large-scale pretraining on paired video-text data enables robust cross-modal alignment, facilitating both retrieval and video question answering. Indeed, models like VideoCLIP and VIOLET also demonstrate strong QA performance (*e.g.*, 92.1% on MSVD-QA, 68.9% on TGIF-QA), showing that semantic transfer extends beyond simple retrieval. Nevertheless, these models generally report limited or no performance on temporal localization benchmarks, suggesting that video-language alignment alone cannot fully replace specialized temporal reasoning.

The most pronounced performance leap occurs with large-scale models such as InternVideo and InternVideo2. These models achieve state-of-the-art retrieval results across multiple datasets (MSR-VTT: 62.8%, ActivityNet: 74.1%, VATEX: 75.5%) while maintaining strong localization performance (THUMOS'14 mAP: 72.0%). They also exhibit high accuracy on multiple-choice QA (MSR-VTT: 93.4%, LSMDC: 77.3%), highlighting their ability to generalize multimodal understanding to structured reasoning tasks. These results demonstrate that scaling both model size and pretraining diversity enhances not only cross-modal alignment but also downstream adaptability across datasets of varying complexity and granularity. Instruction-tuned video-language models such as Video-ChatGPT, Valley, and Grounding-GPT introduce a complementary paradigm. By aligning video understanding with natural language instructions, these models excel in video QA, achieving top-1 accuracies exceeding 49% on MSR-VTT-QA and TGIF-QA, despite minimal supervised fine-tuning. Their performance highlights the potential of instruction tuning to enable flexible, open-ended reasoning, although retrieval metrics remain largely unreported.

Across datasets, several trends emerge. Short-clip retrieval datasets like MSR-VTT and MSVD benefit most from large-scale video-language pretraining, whereas long-range, compositional datasets such as ActivityNet and DiDeMo show the importance of temporal reasoning. QA datasets, including MSVD-QA and TGIF-QA, require both semantic alignment and multimodal reasoning, favoring instruction-tuned architectures. Multiple-choice settings further emphasize the benefits

TABLE IV: Comparison of representative video models across tasks. Dataset abbreviations are as follows: Video retrieval (MSR=MSR-VTT, MSVD=MSVD, LSMDC=LSMDC, ANet=ActivityNet, DiDeMo=DiDeMo, VATEX=VATEX), Temporal action localization (TH'14=THUMOS'14, ANet=ActivityNet, HACS=HACS, FineAct=FineAction), Video question answering (MSRVTT-QA, MSVD-QA, TGIF-QA, ANet-QA), and Multiple-choice (MSR-VTT, LSMDC). Reported metrics: Recall@1 (R@1) for retrieval, average mAP for localization, top-1 accuracy for QA, and zero-shot performance for multiple-choice. Early backbones excel at localization but have limited reported performance in retrieval and QA; video-language pretrained models show strong retrieval and QA; instruction-tuned models (Video-ChatGPT, Grounding-GPT, Valley) excel at zero-shot QA.

Model	Video retrieval						Action localization				Video question answering				Multiple-choice	
	MSR	MSVD	LSMDC	ANet	DiDeMo	VATEX	TH'14	ANet	HACS	FineAct	MSRVTT	MSVD	TGIF	ANet	MSR	LSMDC
I3D [24]+Flow	-	-	-	-	-	-	66.8	35.6	-	-	-	-	-	-	-	-
R(2+1)D [199]	-	-	-	-	-	-	55.6	36.6	-	-	-	-	-	-	-	-
SlowFast [3]	-	-	-	-	-	-	-	-	38.7	-	-	-	-	-	-	-
Heterogeneous [241]	-	-	-	-	-	-	-	-	-	-	33.0	33.7	53.8	-	-	-
VideoCLIP [224]	30.9	-	-	-	-	-	-	-	-	-	92.1	-	-	-	-	-
ClipBERT [242]	22.0	-	-	21.3	20.4	-	-	-	-	-	37.4	-	60.3	-	88.2	-
VIOLET [226]	34.5	-	16.1	-	32.6	-	-	-	-	-	43.9	47.9	68.9	-	-	82.9
Frozen [172]	32.5	33.7	15.0	28.8	34.6	-	-	-	-	-	-	-	-	-	-	-
CLIP4Clip [243]	42.1	46.2	22.6	40.5	43.4	-	-	-	-	-	-	-	-	-	-	-
FrozenBiLM [244]	-	-	-	-	-	-	-	-	-	-	16.8	32.2	41.0	24.7	-	-
ALPRO [227]	33.9	-	-	-	35.9	-	-	-	-	-	42.1	45.9	-	-	-	-
InternVideo [30]	55.2	58.4	34.0	62.2	57.9	71.1	71.6	39.0	41.6	17.6	47.1	55.5	72.2	-	93.4	77.3
VideoMAE V2 [31]	-	-	-	-	-	-	69.6	-	-	18.2	-	-	-	-	-	-
All-in-one [245]	37.3	-	-	22.4	32.7	-	-	-	-	-	46.8	48.3	67.3	-	91.9	83.9
Video Chat [246]	-	-	-	-	-	-	-	-	-	-	45.0	56.3	34.4	26.5	-	-
LLaMA Adapter [247]	-	-	-	-	-	-	-	-	-	-	43.8	54.9	-	34.2	-	-
Video LLaMA [228]	-	-	-	-	-	-	-	-	-	-	29.6	51.6	-	12.4	-	-
Video-ChatGPT [229]	-	-	-	-	-	-	-	-	-	-	49.3	64.9	51.4	35.2	-	-
Valley [248]	-	-	-	-	-	-	-	-	-	-	50.8	69.2	-	44.9	-	-
InternVideo2 [32]	62.8	61.4	46.4	74.1	74.2	75.5	72.0	41.2	43.3	27.7	-	-	-	-	-	-
Grounding-GPT [249]	-	-	-	-	-	-	-	-	-	-	51.6	67.8	-	44.7	-	-

of models that can integrate retrieval, localization, and reasoning capabilities. Taken together, these patterns illustrate a clear trajectory: from modality-specific backbones optimized for temporal cues, through multimodal alignment via video-language pretraining, to instruction-tuned models capable of flexible zero-shot reasoning. Future video models will likely need to merge the temporal precision of early architectures with the broad multimodal and reasoning capabilities of large-scale, instruction-aligned systems, advancing toward truly general-purpose video understanding.

V. A Dataset-Centric Roadmap for Video Understanding

We now provide a prescriptive roadmap showing how dataset properties shape architectures, guiding model selection while balancing scalability and deployment.

A. Dataset Limitations and Future Outlook

Dataset limitations. Despite their pivotal role in shaping model architectures, existing datasets remain constrained by structural limitations that directly influence what models learn and how they generalize. A first challenge is dataset bias. Many benchmarks reflect narrow cultural or environmental contexts, *e.g.*, sports, kitchens, or scripted movies, leading to strong priors that models can exploit without acquiring robust spatiotemporal reasoning skills. Architectures trained on such datasets may achieve high benchmark accuracy yet falter in real-world deployments, where actions, objects, and environments differ markedly from the training distribution. Bias and imbalance in class coverage further skew learning dynamics, amplifying context-specific shortcuts rather than transferable representations. A second limitation lies in annotation cost and granularity. Fine-grained temporal labels, hierarchical task decompositions, and multimodal alignments are expensive and time-consuming to obtain. Consequently, many datasets provide only sparse or weak supervision, with limited temporal density or noisy boundaries. These constraints have architectural consequences: models trained under such supervision often overfit to annotated segments while ignoring unlabelled structure, motivating the rise of weakly supervised, self-supervised, and semi-automatic approaches that attempt to compensate for annotation sparsity. Third, many datasets lack ecological validity. Curated short clips and trimmed action boundaries capture isolated moments rather than continuous, overlapping, and ambiguous workflows that typify everyday activity. As a result, architectures optimized for such curated data, such as clip-based 3D CNNs or trimmed-sequence transformers, struggle when confronted with egocentric videos, multi-agent dynamics, or long-horizon reasoning tasks. The gap between benchmark data and real-world complexity has fueled interest in architectures that incorporate memory, relational reasoning, causal inference, and multimodal grounding to cope with unconstrained environments. Finally, evaluation remains fragmented across datasets, with heterogeneous metrics and inconsistent protocols that make it difficult to assess generalization (Tables III and IV). Models are often fine-tuned to maximize benchmark-specific accuracy or mean average precision, rather than being evaluated for broader capabilities such as compositional reasoning, causal inference, or robustness to distributional shift. This fragmentation not only obscures comparative progress but also shapes architectural incentives, leading to models tuned for leaderboard performance rather than general-purpose understanding.

Future outlook. These limitations point directly to the requirements of next-generation datasets and models. Future benchmarks should move beyond static, domain-specific cor-

pora to embrace diversity, ecological validity, and scalability. Diversity involves curating datasets that capture a broad range of cultures, environments, and activity types, helping to reduce biases that limit generalization. Ecological validity requires capturing continuous, untrimmed, multimodal, and multi-agent activities, enabling models to reason over overlapping workflows and dynamic social contexts. Scalability demands annotation strategies that combine automated labeling, crowdsourcing, and self-supervised alignment, ensuring that datasets can grow without prohibitive human cost. Crucially, future datasets should be explicitly designed to support not only recognition and localization but also higher-level reasoning tasks such as forecasting, causal analysis, and interactive decision-making. Architectures should co-evolve to meet these demands. Long-horizon reasoning will require models with structured memory, hierarchical temporal abstractions, and recurrent attention mechanisms capable of spanning minutes or even hours of activity. Relational and causal modeling will benefit from graph-based and neuro-symbolic hybrids that can disentangle inter-agent dependencies and infer cause-effect relationships. Multimodal grounding will necessitate foundation models that seamlessly integrate visual, auditory, textual, and sensor streams, with mechanisms for balancing modality dominance and coping with misalignment. Moreover, continual and adaptive learning will become essential as datasets increasingly reflect dynamic, open-world conditions where models should adapt to new tasks, domains, and modalities without catastrophic forgetting.

Alongside dataset and architectural innovation, benchmarking practices should also evolve. Standardized cross-dataset protocols can provide a measure of true generalization, while new evaluation metrics should capture dimensions such as compositional generalization, reasoning faithfulness, robustness under noise, and computational efficiency. Open benchmarks that test causal inference, multi-step prediction, and cross-modal reasoning will better reflect the capabilities demanded by real-world video understanding systems.

Taken together, these findings highlight the mutual relationship between datasets and architectures: current benchmark limitations expose model blind spots, while future model requirements drive the need for more representative, scalable, and challenging datasets. Bridging this gap will require community-wide collaboration in dataset curation, annotation, benchmarking, and model design. If pursued systematically, this agenda has the potential to transform video understanding from narrow task performance toward general-purpose, robust, and socially responsible systems, closing the loop between data, inductive bias, and architectural evolution.

B. Datasets as Engines of Architectural Innovation

Datasets are not passive benchmarks; they are the principal structural force shaping model design. As summarized in Table II and reflected in performance trends in Tables III and IV, every major architectural transition in video understanding has been catalyzed by properties encoded in the data: short, trimmed motion corpora favored two-stream CNNs and early 3D ConvNets; long-range sequential datasets demanded temporal aggregation and memory; and multimodal, text-paired corpora precipitated cross-modal transformers and video-language pretraining. In this sense, datasets operate as inductive-bias generators that determine which invariances, *e.g.*, temporal, relational, and semantic, models should internalize to succeed.

Motion complexity sets the limits of generalization. Coarse, high-amplitude datasets rewarded architectures that capture instantaneous motion (*e.g.*, optical-flow streams and shallow 3D filters), but those same inductive biases often fail in cluttered or low-amplitude regimes. FineGym, Diving48, and AVA illustrate the opposite pressures: fine-grained micro-motions, multi-agent interactions, and sparse cues force models toward multi-scale temporal hierarchies, pose/part reasoning, and attention mechanisms that can isolate salient sub-trajectories. The practical implication is clear: training exclusively on coarse-action datasets leads to fragile transfer performance. Motion granularity should be present in the data if we expect robustness in nuanced real-world settings. Compositional and hierarchical structure unlocks procedural reasoning. Instructional datasets such as Breakfast, YouCook2, COIN, and large weakly aligned corpora like HowTo100M show that activities are not atomic labels but sequences of sub-actions arranged taxonomically and contextually. Exposure to such structure allows models to learn reusable primitives and combine them zero-shot into unseen activities, an ability crucial for robotics, assistive AI, and instructional video analysis. Table II makes this visible in the *Step/Hier* annotation columns: when the dataset records relations among actions rather than only their names, models learn mechanisms that transfer.

Temporal richness and multimodal alignment provide the foundation for deployable systems. Long-horizon corpora and egocentric datasets (*e.g.*, Charades, EPIC-KITCHENS, Ego4D) compel models to track extended dependencies, handle overlapping actions, and disentangle ego-motion. When videos are paired with language, audio, and narration (*e.g.*, ActivityNet Captions, HowTo100M), models are pressured to ground perception in text and sound, enabling cross-modal retrieval and zero-shot generalization. The performance patterns in Table IV mirror this: video-language pretraining and instruction tuning lift retrieval and QA substantially, while pure visual pretraining alone is insufficient for semantic grounding.

These observations also expose concrete dataset gaps. Web-scale video-text corpora deliver breadth, yet their captions are noisy, weakly aligned in time, and culturally/language biased; fine-grained manipulation and low-amplitude motions are underrepresented; spatio-temporal boxes remain sparse outside a handful of benchmarks; and true long-horizon, multi-agent, multimodal datasets with dense alignment are rare. Cross-view and ego-exo bridging are inconsistently available, and compositional generalization is seldom stress-tested with principled splits. Addressing these gaps requires treating dataset design as a strategic lever rather than a scaling exercise. Simply enlarging class vocabularies or clip counts will not yield general video intelligence. The decisive ingredient is *structure*: motion granularity, procedural and hierarchical annotations, temporal continuity across minutes or hours, precise audio/text alignment at sub-second resolution, and evaluation splits that

diagnose composition and transfer. Datasets designed with these properties have historically driven the architectural innovations that underpin the field today; future corpora should be crafted to sustain this virtuous cycle.

C. Open Challenges and Future Directions

The patterns observed in Tables III and IV illustrate a clear principle: datasets generate invariance pressures, which architectures evolve to accommodate. Trimmed, homogeneous clips permitted two-stream and 3D CNNs to dominate short-clip recognition; longer, untrimmed activities exposed their rigidity and motivated temporal convolutions and recurrent aggregation; long-horizon reasoning elevated transformers with scalable attention; rich human-object and multi-agent interactions stimulated relational encoders and graph-augmented models; and, finally, multimodal corpora forced fusion modules and large-scale video-language pretraining. The prescription that follows is not a linear "replace the old with the new", but a matching of inductive bias to dataset structure, together with a plan to close the remaining gaps.

For short and relatively homogeneous data, compact 3D CNNs and hybrid CNN-transformer encoders remain efficient and competitive, particularly when deployment constraints prioritize throughput and latency. As temporal span grows and events occur concurrently, attention mechanisms with memory compression, segment-level pooling, and hierarchical temporal pyramids become essential to preserve long-range context without sacrificing resolution at action boundaries. Where multi-agent interactions and human-object relations are central, relational modules and graph-enhanced transformers provide the inductive bias to track entities, roles, and contact dynamics over time. When audio and language are present, alignment losses and contrastive or generative video-language pretraining become the standard framework for retrieval, captioning, and question answering, as reflected in the substantial gains of CLIP-style and instruction-tuned models.

However, two system-level gaps remain. The first is *temporal-semantic unification*: models that achieve precise temporal localization often lack open-vocabulary semantic understanding, whereas models with strong semantic capabilities (*e.g.*, retrieval or QA) perform poorly at boundary-level localization. A promising direction is to couple dense temporal detectors with token-level audio-text grounding, sharing representations across localization and language understanding heads. The second is *long-horizon compositional reasoning*: current instruction-tuned systems excel at zero-shot QA but degrade on extended, multi-step procedures. Here, retrieval-augmented video understanding, indexing events and steps into a persistent memory and querying them with language, offers a practical path forward, especially when paired with datasets that provide chaptering, steps, and cross-modal timestamps. Progress also depends on evaluation that measures what matters. Alongside top-1 and R@1, reporting should emphasize temporal mAP across IoU thresholds for localization, moment retrieval under compositional splits, QA accuracy under counterfactual and long-horizon subsets, calibration (*e.g.*, ECE/Brier) for safety-critical use, and compute/latency metrics for deployment. Cross-dataset testing, ego to exo, lab to in-the-wild, language and culture shifts, should become a first-class protocol rather than an afterthought. These practices are directly tied to the contributions: we implement a dataset-centric perspective, relate structural properties to observed performance trends across recognition, localization, retrieval, and QA, and convert this synthesis into concrete guidance for selecting and designing models under real-world constraints.

The roadmap is therefore dual. On the modeling side, pursue architectures that integrate the temporal precision of CNNs, the hierarchical composition needed for procedures, the scalability of transformers for long horizons, and the grounding provided by multimodal pretraining and instruction tuning. On the data side, build the corpora that will pressure such models to emerge: long-form, multi-agent, multimodal datasets with sub-second alignment; annotations that expose steps, roles, and relations; splits that test compositional generalization and domain transfer; and curated hard negatives that probe fine-grained motion and language disambiguation. If dataset and model co-evolve along these lines, the field can move beyond recognition of isolated clips toward robust, general, and deployable video understanding, precisely the dataset-driven vision advanced by this survey.

VI. Conclusion

Video understanding has evolved from short-clip recognition to a foundation for multimodal, relational, and long-horizon reasoning. In this survey, we adopt a dataset-driven perspective, showing how structural properties have shaped architectural evolution. We show clear patterns: 3D CNNs and two-stream networks excel on short, motion-focused clips; sequential and transformer models thrive on compositional and procedural tasks; and large-scale, instruction-tuned video-language models succeed when semantic grounding and cross-modal reasoning are required. Looking forward, general-purpose video understanding will arise from the co-evolution of models and datasets: architectures should integrate temporal precision, hierarchical reasoning, and multimodal grounding, while datasets should expose fine-grained motions, procedural hierarchies, multi-agent interactions, and dense cross-modal alignment. By emphasizing structure over scale, the field can move beyond isolated recognition toward robust, context-aware, and deployable video intelligence, with datasets as the central drivers of progress.

References

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[4] Q. Chen, L. Wang, P. Koniusz, and T. Gedeon, "Motion meets attention: Video motion prompts," in *Asian Conference on Machine Learning*. PMLR, 2025, pp. 591–606.

[5] L. Wang, X. Yuan, T. Gedeon, and L. Zheng, "Taylor videos for action recognition," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 52 117–52 133.

[6] D. Ding, L. Wang, L. Zhu, T. Gedeon, and P. Koniusz, "Learnable expansion of graph operators for multi-modal feature fusion," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=SMZqIOSdlN

[7] L. Wang, J. Liu, L. Zheng, T. Gedeon, and P. Koniusz, "Meet jeanie: a similarity measure for 3d skeleton sequences via temporal-viewpoint alignment," *International Journal of Computer Vision*, vol. 132, no. 9, pp. 4091–4122, 2024.

[8] J. Qiu and L. Wang, "Evolving skeletons: Motion dynamics in action recognition," in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 1916–1937.

[9] L. Wang and P. Koniusz, "Feature hallucination for self-supervised action recognition," *International Journal of Computer Vision*, 2025.

[10] X. Ding and L. Wang, "Do language models understand time?" in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 1855–1868.

[11] ——, "Quo vadis, anomaly detection? llms and vlms in the spotlight," *arXiv preprint arXiv:2412.18298*, 2024.

[12] ——, "The journey of action recognition," in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 1869–1884.

[13] Y. Liu, J. Yang, M. Perera, P. Ji, D. Kim, M. Xu, T. Wang, S. Anwar, T. Gedeon, L. Wang *et al.*, "Representation-centric survey of skeletal action recognition and the anubis benchmark," *CoRR*, 2025.

[14] N. Madan, A. Møgelmose, R. Modi, Y. S. Rawat, and T. B. Moeslund, "Foundation models for video understanding: A survey," *arXiv preprint arXiv:2405.03770*, 2024.

[15] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu *et al.*, "Video understanding with large language models: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[16] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[17] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. 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*, 2017, pp. 5842–5850.

[18] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[19] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European conference on computer vision*. Springer, 2016, pp. 510–526.

[20] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," *arXiv preprint arXiv:1804.09626*, 2018.

[21] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6047–6056.

[22] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision*, vol. 130, no. 1, pp. 33–55, 2022.

[23] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang, "Graph based skeleton motion representation and similarity measurement for action recognition," in *European conference on computer vision*. Springer, 2016, pp. 370–385.

[24] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[26] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[27] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[28] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Icml*, vol. 2, no. 3, 2021, p. 4.

[29] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[30] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang *et al.*, "Internvideo: General video foundation models via generative and discriminative learning," *arXiv preprint arXiv:2212.03191*, 2022.

[31] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 549–14 560.

[32] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi *et al.*, "Internvideo2: Scaling foundation models for multimodal video understanding," in *European Conference on Computer Vision*. Springer, 2024, pp. 396–416.

[33] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[34] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[35] A. Klaser, M. Marsza lek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British machine vision conference*. British Machine Vision Association, 2008, pp. 275–1.

[36] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[37] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 720–736.

[38] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[39] J. Wang, D. Chen, Z. Wu, C. Luo, L. Zhou, Y. Zhao, Y. Xie, C. Liu, Y.-G. Jiang, and L. Yuan, "Omnivl: One foundation model for image-language and video-language tasks," *Advances in neural information processing systems*, vol. 35, pp. 5696–5710, 2022.

[40] J. Wu, M. Zhong, S. Xing, Z. Lai, Z. Liu, Z. Chen, W. Wang, X. Zhu, L. Lu, T. Lu *et al.*, "Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 69 925–69 975, 2024.

[41] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou, "Timechat: A time-sensitive multimodal large language model for long video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 313–14 323.

[42] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li *et al.*, "Videollama 3: Frontier multimodal foundation models for image and video understanding," *arXiv preprint arXiv:2501.13106*, 2025.

[43] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi, "Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 439–26 455.

[44] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 668–14 678.

[45] X. Xu, T. Hospedales, and S. Gong, "Transductive zero-shot action recognition by word-vector embedding," *International Journal of Computer Vision*, vol. 123, no. 3, pp. 309–333, 2017.

[46] D. Mandal, S. Narayan, S. K. Dwivedi, V. Gupta, S. Ahmed, F. S. Khan, and L. Shao, "Out-of-distribution detection for generalized zero-shot action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9985–9993.

[47] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *European conference on computer vision*. Springer, 2020, pp. 525–542.

[48] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[49] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen, "Temporal-relational crosstransformers for few-shot action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 475–484.

[50] L. Wang and P. Koniusz, "Temporal-viewpoint transportation plan for skeletal few-shot action recognition," in *Proceedings of the Asian conference on computer vision*, 2022, pp. 4176–4193.

[51] ——, "Uncertainty-dtw for time series and sequences," in *European Conference on Computer Vision*. Springer, 2022, pp. 176–195.

[52] R. Pasunuru and M. Bansal, "Reinforced video captioning with entailment rewards," *arXiv preprint arXiv:1708.02300*, 2017.

[53] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1–10.

[54] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in neural information processing systems*, vol. 34, pp. 23 634–23 651, 2021.

[55] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[56] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3200–3225, 2022.

[57] T. Nguyen, Y. Bin, J. Xiao, L. Qu, Y. Li, J. Z. Wu, C.-D. Nguyen, S.-K. Ng, and L. A. Tuan, "Video-language understanding: A survey from model architecture, model training, and data perspectives," *arXiv preprint arXiv:2406.05615*, 2024.

[58] J. P.J. and B. C. Kovoor, "Video question answering: A survey of the state-of-the-art," *J. Vis. Comun. Image Represent.*, vol. 105, no. C, Dec. 2024. [Online]. Available: https://doi.org/10.1016/j.jvcir.2024.104320

[59] P. Zhou, L. Wang, Z. Liu, Y. Hao, P. Hui, S. Tarkoma, and J. Kangasharju, "A survey on generative ai and llm for video generation, understanding, and streaming," *arXiv preprint arXiv:2404.16038*, 2024.

[60] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *Acm Computing Surveys (Csur)*, vol. 43, no. 3, pp. 1–43, 2011.

[61] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[62] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021.

[63] A. Ulhaq, N. Akhtar, G. Pogrebna, and A. Mian, "Vision transformers for action recognition: A survey," *arXiv preprint arXiv:2209.05700*, 2022.

[64] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 922–12 943, 2023.

[65] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9879–9889.

[66] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[67] T. Subetha and S. Chitrakala, "A survey on human activity recognition from videos," in *2016 international conference on information communication and embedded systems (ICICES)*. IEEE, 2016, pp. 1–7.

[68] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *IEEE Transactions on Image Processing*, vol. 29, pp. 15–28, 2019.

[69] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 183–192.

[70] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph convolutional neural network for human action recognition: A comprehensive survey," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 128–145, 2021.

[71] L. Mourot, L. Hoyet, F. Le Clerc, F. Schnitzler, and P. Hellier, "A survey on deep learning for skeleton-based human animation," in *Computer Graphics Forum*, vol. 41, no. 1. Wiley Online Library, 2022, pp. 122–157.

[72] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *Cyborg and Bionic Systems*, vol. 5, p. 0100, 2024.

[73] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.

[74] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.

[75] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, 2022.

[76] A. Oussidi and A. Elhassouny, "Deep generative models: Survey," in *2018 International conference on intelligent systems and computer vision (ISCV)*. IEEE, 2018, pp. 1–8.

[77] M. Suzuki and Y. Matsuo, "A survey of multimodal deep generative models," *Advanced Robotics*, vol. 36, no. 5-6, pp. 261–278, 2022.

[78] J. Cho, F. D. Puspitasari, S. Zheng, J. Zheng, L.-H. Lee, T.-H. Kim, C. S. Hong, and C. Zhang, "Sora as an agi world model? a complete survey on text-to-video generation," *arXiv preprint arXiv:2403.05131*, 2024.

[79] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, "A survey on video diffusion models," *ACM Computing Surveys*, vol. 57, no. 2, pp. 1–42, 2024.

[80] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli, "Benchmarking a multimodal and multiview and interactive dataset for human action recognition," *IEEE Transactions on cybernetics*, vol. 47, no. 7, pp. 1781–1794, 2016.

[81] T. Singh and D. K. Vishwakarma, "Video benchmarks of human action datasets: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1107–1154, 2019.

[82] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," *Pattern Recognition Letters*, vol. 118, pp. 14–22, 2019.

[83] D. Guo, K. Li, B. Hu, Y. Zhang, and M. Wang, "Benchmarking micro-action recognition: Dataset, methods, and applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6238–6252, 2024.

[84] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G. M. Farinella, D. Damen, and T. Tommasi, "An outlook into the future of egocentric vision," *International Journal of Computer Vision*, vol. 132, no. 11, pp. 4880–4936, 2024.

[85] H.-C. Shih, "A survey of content-aware video analysis for sports," *IEEE Transactions on circuits and systems for video technology*, vol. 28, no. 5, pp. 1212–1231, 2017.

[86] I. Ahmad, X. Wei, Y. Sun, and Y.-Q. Zhang, "Video transcoding: an overview of various techniques and research issues," *IEEE Transactions on multimedia*, vol. 7, no. 5, pp. 793–804, 2005.

[87] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–37, 2019.

[88] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6212–6221.

[89] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[90] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: a review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–25, 2022.

[91] M. Zhao, Y. Yu, X. Wang, L. Yang, and D. Niu, "Search-map-search: a frame selection paradigm for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 627–10 636.

[92] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–37, 2023.

[93] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2, pp. 107–123, 2005.

[94] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[95] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[96] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[97] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[98] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[99] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural networks*, vol. 113, pp. 54–71, 2019.

[100] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[101] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[102] X. Yin, Y. Zhu, and J. Hu, "A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–36, 2021.

[103] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36.

[104] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1395–1402.

[105] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[106] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[107] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 2929–2936.

[108] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*. IEEE, 2009, pp. 1282–1289.

[109] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *European conference on computer vision*. Springer, 2010, pp. 392–405.

[110] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 9–14.

[111] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2012, pp. 20–27.

[112] V. Bloom, D. Makris, and V. Argyriou, "G3d: A gaming action dataset and real time action recognition evaluation framework," in *2012 IEEE Computer society conference on computer vision and pattern recognition workshops*. IEEE, 2012, pp. 7–12.

[113] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine vision and applications*, vol. 24, no. 5, pp. 971–981, 2013.

[114] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 479–485.

[115] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[116] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2649–2656.

[117] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[118] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann, "Infar dataset: Infrared action recognition at different times," *Neurocomputing*, vol. 212, pp. 36–47, 2016.

[119] S. Vadivelu, S. Ganesan, O. R. Murthy, and A. Dhall, "Thermal imaging based elderly fall detection," in *Asian conference on computer vision*. Springer, 2016, pp. 541–553.

[120] P. Weinzaepfel, X. Martin, and C. Schmid, "Human action localization with sparse spatial supervision," *arXiv preprint arXiv:1605.05197*, 2016.

[121] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, "Every moment counts: Dense detailed labeling of actions in complex videos," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 375–389, 2018.

[122] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1971–1980.

[123] H. Kiani Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey, "Need for speed: A benchmark for higher frame rate object tracking," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1125–1134.

[124] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[125] G. Kanojia, S. Kumawat, and S. Raman, "Attentive spatio-temporal representation learning for diving classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[126] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick *et al.*, "Moments in time dataset: one million videos for event understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 502–508, 2019.

[127] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[128] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2616–2625.

[129] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[130] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "Ava active speaker: An audio-visual dataset for active speaker detection," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 4492–4496.

[131] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 266–16 275.

[132] A. A. Gritsenko, X. Xiong, J. Djolonga, M. Dehghani, C. Sun, M. Lucic, C. Schmid, and A. Arnab, "End-to-end spatio-temporal action localisation with video transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 373–18 383.

[133] J. Huh, J. Chalk, E. Kazakos, D. Damen, and A. Zisserman, "Epic-sounds: A large-scale dataset of actions that sound," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[134] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *2013 IEEE workshop on applications of computer vision (WACV)*. IEEE, 2013, pp. 53–60.

[135] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 842–849.

[136] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision*. Springer, 2012, pp. 314–327.

[137] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International journal of robotics research*, vol. 32, no. 8, pp. 951–970, 2013.

[138] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 729–738.

[139] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 780–787.

[140] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 287–295.

[141] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5344–5352.

[142] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[143] Y. Li, M. Liu, and J. 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)*, 2018, pp. 619–635.

[144] L. Zhou, N. Louis, and J. J. Corso, "Weakly-supervised video object grounding from text by loss weighting and object interaction," *arXiv preprint arXiv:1805.02834*, 2018.

[145] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "Coin: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1207–1216.

[146] R. Girdhar and D. Ramanan, "Cater: A diagnostic dataset for compositional actions and temporal reasoning," *arXiv preprint arXiv:1910.04744*, 2019.

[147] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," *arXiv preprint arXiv:1910.01442*, 2019.

[148] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3537–3545.

[149] Z. Luo, W. Xie, S. Kapoor, Y. Liang, M. Cooper, J. C. Niebles, E. Adeli, and F.-F. Li, "Moma: Multi-object multi-actor activity parsing," *Advances in neural information processing systems*, vol. 34, pp. 17 939–17 955, 2021.

[150] Z. Luo, Z. Durante, L. Li, W. Xie, R. Liu, E. Jin, Z. Huang, L. Y. Li, J. Wu, J. C. Niebles *et al.*, "Moma-lrg: Language-refined graphs for multi-object multi-actor activity parsing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5282–5298, 2022.

[151] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2847–2854.

[152] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1290–1297.

[153] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1194–1201.

[154] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *Computer Vision and Image Understanding*, vol. 155, pp. 1–23, 2017.

[155] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.

[156] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1711–1721.

[157] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "Hacs: Human action clips and segments dataset for recognition and temporal localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8668–8678.

[158] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action analytics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–24, 2020.

[159] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[160] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[161] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[162] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 94–120, 2017.

[163] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5803–5812.

[164] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[165] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2758–2766.

[166] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1645–1653.

[167] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5267–5275.

[168] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," *arXiv preprint arXiv:1809.01696*, 2018.

[169] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4581–4591.

[170] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9777–9786.

[171] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "Agqa: A benchmark for compositional spatio-temporal reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 287–11 297.

[172] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738.

[173] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5036–5045.

[174] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. 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*, 2022, pp. 18 995–19 012.

[175] A. Yang, A. Nagrani, I. Laptev, J. Sivic, and C. Schmid, "Vidchapters-7m: Video chapters at scale," *Advances in Neural Information Processing Systems*, vol. 36, pp. 49 428–49 444, 2023.

[176] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang *et al.*, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2307.06942*, 2023.

[177] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-w. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang *et al.*, "Panda-70m: Captioning 70m videos with multiple cross-modality teachers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 320–13 331.

[178] X. Ju, Y. Gao, Z. Zhang, Z. Yuan, X. Wang, A. Zeng, Y. Xiong, Q. Xu, and Y. Shan, "Miradata: A large-scale video dataset with long durations and structured captions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 48 955–48 970, 2024.

[179] K. Nan, R. Xie, P. Zhou, T. Fan, Z. Yang, Z. Chen, X. Li, J. Yang, and Y. Tai, "Openvid-1m: A large-scale high-quality dataset for text-to-video generation," *arXiv preprint arXiv:2407.02371*, 2024.

[180] Q. Wang, Y. Shi, J. Ou, R. Chen, K. Lin, J. Wang, B. Jiang, H. Yang, M. Zheng, X. Tao *et al.*, "Koala-36m: A large-scale video dataset

improving consistency between fine-grained conditions and video content," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8428–8437.

[181] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The ava-kinetics localized human actions video dataset," *arXiv preprint arXiv:2005.00214*, 2020.

[182] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[183] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. 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*, 2019, pp. 2630–2640.

[184] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[185] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[186] R. Girdhar and K. Grauman, "Anticipative video transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 505–13 515.

[187] H. Wang, D. Tran, L. Torresani, and M. Feiszli, "Video modeling with correlation networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 352–361.

[188] S. Sudhakaran, S. Escalera, and O. Lanz, "Gate-shift networks for video action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1102–1111.

[189] X. Chen, A. Pang, W. Yang, Y. Ma, L. Xu, and J. Yu, "Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos," *International Journal of Computer Vision*, vol. 129, no. 10, pp. 2846–2864, 2021.

[190] M. Kim, P. H. Seo, C. Schmid, and M. Cho, "Learning correlation structures for vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 18 941–18 951.

[191] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3575–3584.

[192] F. Sener, D. Singhania, and A. Yao, "Temporal aggregate representations for long-range video understanding," in *European conference on computer vision*. Springer, 2020, pp. 154–171.

[193] J. Wang, W. Zhu, P. Wang, X. Yu, L. Liu, M. Omar, and R. Hamid, "Selective structured state-spaces for long-form video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6387–6397.

[194] C. Feichtenhofer, Y. Li, K. He *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in neural information processing systems*, vol. 35, pp. 35 946–35 958, 2022.

[195] R. Girdhar, M. Singh, N. Ravi, L. Van Der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 102–16 112.

[196] C. Ryali, Y.-T. Hu, D. Bolya, C. Wei, H. Fan, P.-Y. Huang, V. Aggarwal, A. Chowdhury, O. Poursaeed, J. Hoffman *et al.*, "Hiera: A hierarchical vision transformer without the bells-and-whistles," in *International conference on machine learning*. PMLR, 2023, pp. 29 441–29 454.

[197] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[198] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[199] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[200] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 284–293.

[201] H. Wu, Y. Chen, N. Wang, and Z. Zhang, "Sequence level semantics aggregation for video object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9217–9225.

[202] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5492–5501.

[203] N. Hussein, E. Gavves, and A. W. Smeulders, "Timeception for complex action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 254–263.

[204] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7464–7473.

[205] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 122–132.

[206] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 591–600.

[207] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *European Conference on Computer Vision*. Springer, 2020, pp. 214–229.

[208] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in neural information processing systems*, vol. 34, pp. 24 206–24 221, 2021.

[209] C.-Y. Wu, Y. Li, K. Mangalam, H. Fan, B. Xiong, J. Malik, and C. 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*, 2022, pp. 13 587–13 597.

[210] M. M. Islam and G. Bertasius, "Long movie clip classification with state-space video models," in *European Conference on Computer Vision*. Springer, 2022, pp. 87–104.

[211] Z. Wang, M. Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem *et al.*, "Language models with image descriptors are strong few-shot video-language learners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8483–8497, 2022.

[212] D. Liu, Q. Li, A.-D. Dinh, T. Jiang, M. Shah, and C. Xu, "Diffusion action segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10 139–10 149.

[213] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 714–10 726.

[214] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," in *European conference on computer vision*. Springer, 2024, pp. 237–255.

[215] B. He, H. Li, Y. K. Jang, M. Jia, X. Cao, A. Shah, A. Shrivastava, and S.-N. Lim, "Ma-lmm: Memory-augmented large multimodal model for long-term video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 504–13 514.

[216] S. Li, H. Singh, and A. Grover, "Mamba-nd: Selective state space modeling for multi-dimensional data," in *European Conference on Computer Vision*. Springer, 2024, pp. 75–92.

[217] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 244–253.

[218] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 203–213.

[219] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6964–6974.

[220] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, and Y. Qiao, "Unmasked teacher: Towards training-efficient video foundation models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19 948–19 960.

[221] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 803–818.

[222] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 399–417.

[223] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6824–6835.

[224] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," *arXiv preprint arXiv:2109.14084*, 2021.

[225] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu, "Videococa: Video-text modeling with zero-shot transfer from contrastive captioners," *arXiv preprint arXiv:2212.04979*, 2022.

[226] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "Violet: End-to-end video-language transformers with masked visual-token modeling," *arXiv preprint arXiv:2111.12681*, 2021.

[227] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4953–4963.

[228] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[229] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[230] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4597–4605.

[231] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.

[232] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7083–7093.

[233] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Tiny video networks," *Applied AI Letters*, vol. 3, no. 1, p. e38, 2022.

[234] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "Movinets: Mobile video networks for efficient video recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 020–16 030.

[235] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 695–712.

[236] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3163–3172.

[237] M. S. Ryoo, A. Piergiovanni, M. Tan, and A. Angelova, "Assemblenet: Searching for multi-stream neural connectivity in video architectures," *arXiv preprint arXiv:1905.13209*, 2019.

[238] M. Patrick, D. Campbell, Y. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 493–12 506, 2021.

[239] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[240] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4804–4814.

[241] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1999–2007.

[242] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. 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*, 2021, pp. 7331–7341.

[243] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval," *arXiv preprint arXiv:2104.08860*, 2021.

[244] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Zero-shot video question answering via frozen bidirectional language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 124–141, 2022.

[245] J. Wang, Y. Ge, R. Yan, Y. Ge, K. Q. Lin, S. Tsutsui, X. Lin, G. Cai, J. Wu, Y. Shan *et al.*, "All in one: Exploring unified video-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6598–6608.

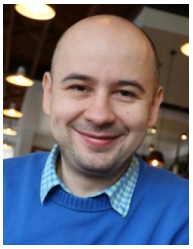
[246] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[247] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.

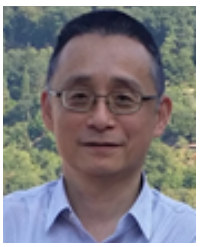
[248] R. Luo, Z. Zhao, M. Yang, J. Dong, D. Li, P. Lu, T. Wang, L. Hu, M. Qiu, and Z. Wei, "Valley: Video assistant with large language model enhanced ability," *arXiv preprint arXiv:2306.07207*, 2023.

[249] Z. Li, Q. Xu, D. Zhang, H. Song, Y. Cai, Q. Qi, R. Zhou, J. Pan, Z. Li, V. T. Vu *et al.*, "Groundinggpt: Language enhanced multi-modal grounding model," *arXiv preprint arXiv:2401.06071*, 2024.

Lei Wang received his M.E. in Software Engineering from the University of Western Australia (UWA) in 2018 and his Ph.D. in Engineering and Computer Science from the Australian National University (ANU) in 2023. He is a Research Fellow in the School of Electrical and Electronic Engineering at Griffith University and a Visiting Scientist with Data61/CSIRO. He leads the Temporal Intelligence and Motion Extraction (TIME) Lab at the ARC Research Hub and Griffith University. He previously held research positions at ANU, UWA, and Data61/CSIRO. His research focuses on motion-, data-, and model-centric approaches to video action recognition and anomaly detection. He has authored numerous first-author papers in top-tier venues, including CVPR, ICCV, ECCV, ACM Multimedia, TPAMI, IJCV, and TIP, and received the Sang Uk Lee Best Student Paper Award at ACCV 2022. He serves as an Area Chair for ACM Multimedia 2024–2025, ICASSP 2025, and ICPR 2024, and was recognized as an Outstanding Area Chair at ACM Multimedia 2024.



Piotr Koniusz received the BSc degree in Telecommunications and Software Engineering from Warsaw University of Technology, Poland, in 2004, and the PhD degree in Computer Vision from CVSSP, University of Surrey, U.K., in 2013. He is a Senior Researcher with the Machine Learning Research Group, Data61/CSIRO, and a Senior Honorary Lecturer at the Australian National University (ANU). He was previously a postdoctoral researcher with the LEAR team at INRIA, France. His research interests include representation learning (contrastive and self-supervised learning), vision-language models, LLMs, and deep and graph neural networks, as well as image classification and action recognition. He has received awards including the Sang Uk Lee Best Student Paper Award (ACCV 2022), Runner-up APRS/IAPR Best Student Paper Award (DICTA 2022), and Outstanding Area Chair recognition (ICLR 2021–2023). He serves as a Workshop Program Chair and Senior Area Chair for NeurIPS 2023 and is Program Chair for NeurIPS 2025.



Yongsheng Gao (Senior Member, IEEE) received the BSc and MSc degrees in Electronic Engineering from Zhejiang University, China, in 1985 and 1988, respectively, and the PhD degree in Computer Engineering from Nanyang Technological University, Singapore. He is currently a Professor at the School of Engineering and Built Environment, Griffith University, and Director of the ARC Research Hub for Driving Farming Productivity and Disease Prevention, Australia. He was previously the Leader of the Biosecurity Group at the Queensland Research Laboratory, National ICT Australia (ARC Centre of Excellence), a consultant at Panasonic Singapore Laboratories, and an Assistant Professor at Nanyang Technological University. His research interests include smart farming, machine vision for agriculture, biosecurity, face recognition, biometrics, image retrieval, computer vision, pattern recognition, environmental informatics, and medical imaging. He is the recipient of the 2025 ARC Industry Laureate Fellowship.