Do Language Models Understand Time?

Xi Ding
Australian National University
Canberra, Australian Capital Territory, Australia
Xi.Ding1@anu.edu.au

Lei Wang*
Griffith University
Brisbane, Queensland, Australia
Australian National University
Canberra, Australian Capital Territory, Australia
l.wang4@griffith.edu.au

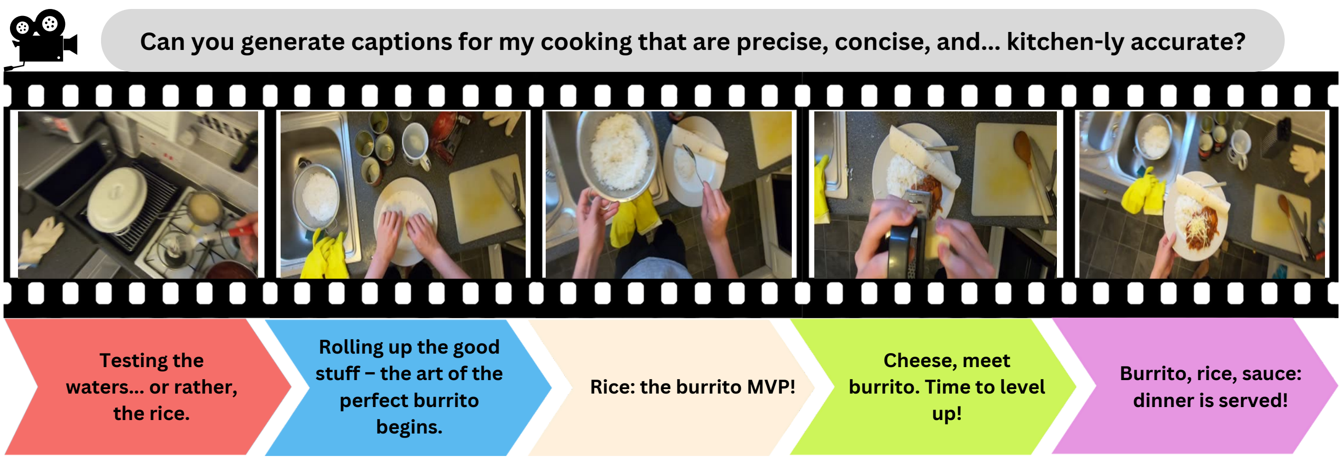


Figure 1: Do language models understand time? In the kitchen arena, where burritos are rolled, rice waits patiently, and sauce steals the spotlight, LLMs try their best to keep up. Captions flow like a recipe—precise and tempting—but can they truly tell the difference between prepping, cooking, and eating? After all, in cooking, timing isn't just everything—it's the secret sauce!

Abstract

Large language models (LLMs) have revolutionized video-based computer vision applications, including action recognition, anomaly detection, and video summarization. Videos inherently pose unique challenges, combining spatial complexity with temporal dynamics that are absent in static images or textual data. Current approaches to video understanding with LLMs often rely on pretrained video encoders to extract spatiotemporal features and text encoders to capture semantic meaning. These representations are integrated within LLM frameworks, enabling multimodal reasoning across diverse video tasks. However, the critical question persists: Can LLMs truly understand the concept of time, and how effectively can they reason about temporal relationships in videos? This work critically examines the role of LLMs in video processing, with a specific focus on their temporal reasoning capabilities. We identify key limitations in the interaction between LLMs and pretrained encoders, revealing gaps in their ability to model long-term dependencies and abstract temporal concepts such as causality and event progression. Furthermore, we analyze challenges posed by existing video datasets, including biases, lack of temporal annotations, and domain-specific limitations that constrain the temporal understanding of LLMs. To address these gaps, we explore promising future directions, including the co-evolution of LLMs and encoders, the development of enriched datasets with explicit temporal labels, and innovative architectures for integrating spatial, temporal, and semantic reasoning. By addressing these challenges, we aim to advance the temporal comprehension of LLMs, unlocking their full potential in video analysis and beyond. Our paper's GitHub repository can be found here.

*Corresponding author.

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717744

CCS Concepts

• **Computing methodologies** → **Motion capture**; *Neural networks*; • **Information systems** → *Language models*; World Wide Web.

Keywords

Language language models, Videos, Temporal, Interaction

ACM Reference Format:
Xi Ding and Lei Wang. 2025. Do Language Models Understand Time?. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3701716.3717744

1 Introduction

Large language models (LLMs) have brought transformative advancements to artificial intelligence (AI), excelling across a wide array of tasks in natural language processing and computer vision [11, 53, 186]. Their ability to understand and generate human-like language has enabled groundbreaking applications, from machine translation to image and video captioning [83] (see Figure 1, frames from EPIC-KITCHENS-100 [39]). More recently, the integration of LLMs into video processing has sparked significant interest, leading to advances in tasks such as action recognition [177, 178], anomaly detection [160, 202, 212], and video summarization [74, 104, 193, 213]. However, videos pose unique challenges compared to other modalities due to their dual reliance on both spatial and temporal information [21]. Unlike static images, videos capture the dimension of time, embedding sequential dynamics that demand sophisticated reasoning [101, 149]. Similarly, unlike textual data, videos involve rich, complex visual elements that require intricate modeling [26, 170].

Despite these advancements, a fundamental question remains unresolved: Do language models truly understand the concept of time? Temporal reasoning, the ability to comprehend and infer relationships between events over time, is essential for video-based tasks such as causal inference [23, 100], event prediction [63, 72], and understanding action progression [157]. While pretrained video encoders provide LLMs with spatiotemporal embeddings and text encoders contribute semantic insights, the fusion of these components often lacks the nuanced understanding of time required for more advanced applications [44]. Current methods rely heavily on pretrained encoders and dataset-specific tuning [21, 26, 121, 170], raising questions about the generalizability and scalability of these approaches. Table 1 provides a summary of recent video-LLMs, detailing the visual encoders they use and their mechanisms for interacting with these encoders.

This work aims to address these gaps and critically examines the role of LLMs in understanding temporal dynamics in video data. We focus on the interplay between visual (image/video) encoders and LLMs, exploring how effectively they bridge the gap between raw spatiotemporal features and high-level temporal reasoning. By tackling these issues, we seek to shed light on the limitations of existing approaches and inspire innovations in LLM-based video understanding. Our **contributions** are threefold:

i. We provide a detailed review of LLM applications in video processing, with a particular focus on their ability to comprehend temporal concepts, highlighting the state of the art and identifying key limitations.
ii. We analyze the shortcomings of existing LLM-based approaches, particularly their reliance on pretrained encoders and the challenges posed by video datasets, such as lack of temporal annotations and biases towards short-term dependencies.
iii. We propose actionable pathways for advancing LLMs' temporal understanding, emphasizing joint training strategies, better dataset design, and improved alignment between spatiotemporal and semantic features.

By addressing these aspects, this paper seeks to advance our understanding of temporal reasoning in LLMs and pave the way for more robust, generalizable, and scalable solutions in video analysis. The insights offered here aim to engage researchers and practitioners alike, highlighting the importance of bridging the gap between static representations and dynamic reasoning in AI systems.

2 Related Work

The application of LLMs to video processing has attracted significant attention, owing to their capacity to bridge visual and textual modalities [173]. In this section, we review related work across four key areas: LLMs for video understanding, pretrained visual encoders, datasets for video understanding, and temporal reasoning in AI systems. We also highlight the distinct contributions of our work compared to prior studies.

LLMs for video understanding. LLMs have shown remarkable versatility in video-related tasks by incorporating multimodal learning frameworks [184]. Notable works, such as Flamingo [5] by DeepMind, integrate visual and textual modalities for tasks like video captioning and video question answering (QA). Flamingo uses a cross-modal attention mechanism to align spatiotemporal video embeddings with text representations, showcasing the potential of LLMs in multimodal fusion. Other models [55, 57, 153, 197], including OmniVL [153] and Florence [197], explore unified architectures that handle images, videos, and text simultaneously, reducing reliance on domain-specific encoders [84, 161–169]. However, these works primarily focus on improving task performance without a deep analysis of how LLMs handle temporal dynamics, leaving their capacity for explicit time reasoning largely unexamined.

While prior studies [7, 43, 64, 76, 141, 154] primarily emphasize task performance, we specifically investigate whether LLMs truly understand the concept of time. Our work explores the interplay between spatiotemporal embeddings and LLM frameworks, providing a deeper analysis of their temporal reasoning capabilities.

Pretrained visual encoders in multimodal learning. The success of LLMs in video understanding often hinges on the use of pretrained visual encoders [31, 51, 99, 174]. Models like CLIP [122], ResNet [68], and Vision Transformers (ViT) [45] are frequently used for spatial feature extraction, while video-specific encoders such as I3D [18], SlowFast [48], TimeSformer [9], and Video Swin Transformer [109] extract spatiotemporal features. These encoders are trained on large-scale datasets like ImageNet [41, 127] and Kinetics [16, 17, 82], enabling them to capture fine-grained features for downstream tasks. Table 2 presents an overview of widely used pretrained visual encoders along with their corresponding training datasets. Figure 2 compares the performance of visual encoders, showcasing image encoders evaluated on ImageNet-1K[41] and video encoders assessed on Kinetics-400 and Something-Something V2[58]. While the modularity of these encoders facilitates efficient system design, their reliance on general-purpose pretrained features poses limitations, particularly in domain-specific tasks and long-term temporal reasoning [84, 121, 161, 165, 166, 169]. Existing works often treat encoders as static components, overlooking the potential benefits of jointly optimizing encoders and LLMs for temporal understanding [133, 151].

Unlike works that use pretrained encoders as black-box components [134], we examine their limitations, including their bias

Model	Venue	Visual encoder	Other modality encoders	Interaction / Fusion mechanism	Description
Flamingo [5]	NeurIPS 2022	Normalizer-Free ResNet[10]	Text: Chinchilla[70]	Perceiver Resampler & Gated XATTN-DENSE	Visual-language model.
LaViLa [207]	CVPR 2022	TimeSformer[9]	Text: 12-layer Transformer	Cross-attention modules	Large-scale language model.
mPLUG-2 [188]	ICML 2023	CLIP ViT-L/14[122]	Text: BERT[42]	Universal layers & cross-attention modules	Modularized multi-modal foundation model.
Vid2Seq [191]	CVPR 2023	CLIP ViT-L/14[122]	Text: T5-Base[124]	Cross-modal attention	Sequence-to-sequence video-language model.
Video-LLaMA [201]	EMNLP 2023	CLIP ViT-G[122]	Text: Vicuna[35], Audio: ImageBind[54]	Aligned via Q-Formers for video and audio	Instruction-tuned multimodal model.
ChatVideo [151]	arXiv 2023	*e.g.*, OmniVL[153], InternVideo[178]	Text: ChatGPT[182], Audio: *e.g.*, Whisper[122]	Tracklet-centric with ChatGPT reasoning	Chat-based video understanding system.
VideoChat [93]	arXiv 2023	EVA-CLIP ViT-G/14 [140]	Text: StableVicuna[37], Audio: Whisper[122]	Q-Former bridges visual features to LLMs for reasoning	Chat-centric model.
VideoLLM [20]	arXiv 2023	*e.g.*, I3D[18], SlowFast [48]	Text: *e.g.*, BERT[42], T5[124]	Semantic translator aligns visual and text encodings	Video sequence modeling using LLMs.
VAST [27]	NeurIPS 2023	EVA-CLIP ViT-G/14[140]	Text: BERT[42], Audio: BEATs[28]	Cross-attention layers	Omni-modality foundational model.
Video-ChatGPT [115]	ACL 2023	CLIP ViT-L/14 [122]	Text: Vicuna-v1.1[103]	Spatiotemporal features projected via linear layer	Integration of vision and language for video understanding.
Valley [112]	arXiv 2023	CLIP ViT-L/14 [122]	Text: StableVicuna[37]	Projection layer	LLM for video assistant tasks.
Macaw-LLM [113]	arXiv 2023	CLIP ViT-B/16 [122]	Text: LLAMA-7B[146], Audio: Whisper[122]	Alignment module unifies multi-modal representations	Multimodal integration using image, audio, and video inputs.
Auto-AD II [65]	CVPR 2023	CLIP ViT-B/32 [122]	Text: BERT[42]	Cross-attention layers	Movie description using vision and language.
Video-LLaVA [98]	arXiv 2023	LanguageBind-Video [211]	Text: Vicuna v1.5[35]	MLP projection layer	Unified visual representation learning for video.
GPT4Video [180]	ACMMM 2023	CLIP ViT-L/14 [122]	Text: LLaMA 2[146]	Transformer-based cross-attention layer	Video understanding with LLM-based reasoning.
LLaMA-VID [96]	ECCV 2023	CLIP ViT-L/14 [122]	Text: Vicuna[35]	Context attention and linear projector	LLaMA-VID for visual-textual alignment in video.
InternVideo2 [177]	ECCV 2023	InternVL-6B[33], VideoMAE V2 [158]	Text: BERT-Large[42], Audio: BEATs[28]	Q-Former aligns multi-modal embeddings	Foundation model for multimodal video understanding.
COSMO [150]	arXiv 2024	CLIP ViT-L/14[122]	Text: OPT-IML[75]/RedPajama[145]Mistral[80]	Gated cross-attention	Contrastive-streamlined multimodal model.
VTimeLLM [72]	CVPR 2024	CLIP ViT-L/14 [122]	Text: Vicuna[35]	Linear layer	Temporal video understanding enhanced with LLMs.
VILA [99]	CVPR 2024	CLIP ViT-L/336px[122]	Text: LLaMA-2-7B/13B[146]	Linear layer	Vision-language model.
Video ReCap [74]	CVPR 2024	TimeSformer [9]	Text: GPT-2[123]	Cross-attention layers	Recursive hierarchical captioning model
OmniViD [152]	CVPR 2024	VideoSwin [109]	Text: BART[91]	MQ-Former	Generative model for universal video understanding.
VTG-LLM [62]	arXiv 2024	EVA-CLIP ViT-G/14[140]	Text: LLaMA-2-7B[146]	Projection layer	Enhanced video temporal grounding.
AutoAD III [66]	CVPR 2024	EVA-CLIP ViT[140]	Text: GPT-3.5-turbo	Shared Q-Former	Video description enhancement with LLMs.
LAVAD [198]	CVPR 2024	BLIP-2 ViT-L/14, ImageBind [54]	Text: Llama-2-13b-chat[146]	Converts video features into textual prompts for LLMs	Training-free video anomaly detection using LLMs.
MA-LMM [67]	CVPR 2024	EVA-CLIP ViT-G/14	Text: Vicuna[35]	A trainable Q-Former	Memory-augmented large multimodal model.
MiniGPT4-Video [6]	arXiv 2024	EVA-CLIP ViT[140]	Text: LLaMA 2[146]	Concatenates visual tokens and projects into LLM space	Video understanding with visual-textual token interleaving.
PLLaVA [190]	arXiv 2024	CLIP ViT-L/14 [122]	Text: LLAMA-7B[146]	MM projector with adaptive pooling	Parameter-free extension for video captioning tasks.
V2Xum-LLaMA [71]	arXiv 2024	CLIP ViT-L/14 [122]	Text: LLaMA 2[146]	Vision adapter	Video summarization using temporal prompt tuning.
VideoChat2 [94]	CVPR 2024	UMT-L[107]	Text: Vicuna[35]	Linear projection	A comprehensive multi-modal video understanding benchmark.
MotionLLM [25]	arXiv 2024	LanguageBind[211], VQ-VAE[204]	Text: Vicuna[35]	Modality translator: Motion / Video translator	Understanding human behaviors from human motions and videos.
VideoGPT+ [114]	arXiv 2024	CLIP ViT-L/14[122], InternVideo2[177]	Text: Phi-3-Mini-3.8B[1]	MLP	Enhanced video understanding.
EmoLLM [194]	arXiv 2024	CLIP ViT-L/14[122]	Text: Vicuna-v1.5[35], Audio: Whisper[122]	Multi-perspective visual projection	Multimodal emotional understanding with improved reasoning.
Holmes-VAD [202]	arXiv 2024	LanguageBind ViT-L/14[211]	Text: LLaMA3-Instruct-70B[4]	Temporal sampler	Multimodal LLM for video anomaly detection.
ShareGPT4Video [24]	arXiv 2024	CLIP ViT-L/14[122]	Text: Mistral-7B-Instruct-v0.2[80]	MLP	Precise and detailed video captions with hierarchical prompts.
Vriptor [193]	arXiv 2024	EVA CLIP ViT-L/14[140]	Text: ST-LLM[105], Audio: Whisper[122]	Scene-level sequential alignment	Vriptor for dense video captioning.
VideoLLaMA 2 [34]	arXiv 2024	CLIP ViT-L/14[122]	Text: LLAMA 1.5[102], Audio: BEATs[28]	Spatial-Temporal Convolution (STC) connector	Advancing spatial-temporal modeling and audio understanding.
VideoLLM-online [22]	CVPR 2024	CLIP ViT-L/14[122]	Text: Llama-2-Chat[146]/Llama-3-Instruct[4]	MLP projector	Online video large language model for streaming video.
Video-CCAM [47]	arXiv 2024	SigLIP-SO400M[199]	Text: Phi-3-4k-instruct[1]/ Yi-1.5-9B-Chat[3]	Cross-attention-based projector	Causal cross-attention masks for short and long videos.
LongVA [205]	arXiv 2024	CLIP ViT-L/336px [122]	Text: Qwen2-Extended[8, 192]	MLP	Long context video understanding.
InternLM-XComposer-2.5[203]	arXiv 2024	CLIP ViT-L/14 [122]	Text: InternLM2-7B[15], Audio: Whisper[122]	MLP	Long-context LVLM supporting ultra-high-resolution video tasks.
VITA [50]	arXiv 2024	InternViT-300M [31–33, 51]	Text: Mixtral-8x7B [81], Audio: Mel Filter Bank	MLP	Open-source interactive multimodal LLM.
Kangaroo [104]	arXiv 2024	EVA-CLIP-L [140]	Text: Llama-3-8B-Instruct[4]	Multi-modal projector	Video-language model supporting long-context video input.
Qwen2-VL [171]	arXiv 2024	CLIP ViT-L/14[122]	Text: Qwen2-7B[8, 192]	Cross-attention modules	Vision-language model for multimodal tasks.
Oryx [108]	arXiv 2024	OryxViT[108, 199]	Text: Qwen2-7B/32B[8, 192]	Cross-attention	Spatial-temporal model for high-resolution understanding.
Video-XL [135]	arXiv 2024	CLIP ViT-L[122]	Text: Qwen2-7B[8, 192]	Visual-language projector	Long-context video understanding model.
SlowFocus [119]	NeurIPS 2024	CLIP ViT-L/14[122]	Text: Vicuna-7B v1.5[208]	Visual adapter (projector layer)	Fine-grained temporal understanding in video LLM.
VideoStudio [110]	ECCV 2024	CLIP ViT-H/14 [122]	Text: CLIP ViT-H/14[122]	Cross-attention modules	Multi-scene video generation.
VideoINSTA [97]	arXiv 2024	CLIP ViT-L/14 [122]	Text: Llama-3-8B-Instruct[4]	Self-reflective spatial-temporal fusion	Zero-shot long video understanding model.
Loong [179]	arXiv 2024	Causal 3D CNN[195]	Text: Standard text tokenizer	Decoder-only autoregressive LLM with causal attention	Autoregressive language models.
TRACE [63]	arXiv 2024	CLIP ViT-L[122]	Text: Mistral-7B[80]	Task-interleaved sequence modeling & Adaptive head-switching	Video temporal grounding via causal event modeling.
Apollo[213]	arXiv 2024	SigLIP-SO400M[199], InternVideo2[177]	Text: Qwen2.5-7B[8, 192]	Perceiver Resampler & Token Integration with Timestamps	Video understanding model.

Table 1: Summary of latest multimodal video-LLMs and their interaction / fusion mechanisms.

Type	Visual encoder	Pretrained dataset
Image	Normalizer-Free ResNet[10]	ImageNet-1K[41]
	CLIP ViT[122]	WebImageText[138]
	EVA-CLIP ViT[140]	LAION-2B[130], COYO-700M[12]
	BLIP-2 ViT[92]	WebImageText[138]
	SigLIP (*e.g.*, ViT)[199]	WebLI[30]
	OryxViT[108, 199]	WebLI[30]
Video	TimeSformer[9]	Kinetics-400[82], Kinetics-600[16]
	I3D[18]	HMDB51[87], UCF101[137], Kinetics-400[82]
	SlowFast[48]	Kinetics-400[82], Kinetics-600[16], Charades[136]
	VideoSwin[109]	ImageNet-21K[127]
	UMT[107]	Kinetics-400[82], AudioSet[52]
	LanguageBind[211]	VIDAL-10M[211]
	VideoMAE V2[158]	Kinetics-400[82], -600[16], -700[17], Something-Something V2[58], AVA[60]
	InternVL (*e.g.*, InternViT-6B) [32]	Hybrid image-text datasets (*e.g.*, LAION-en)[14, 19, 61, 130, 131]
	InternVideo2[177]	Hybrid video datasets (*e.g.*, Kinetics-400[82], InternVid[176])

Table 2: Visual encoders with their pretrained datasets.

toward short-term dependencies and their challenges in generalizing to abstract temporal concepts. We propose pathways for jointly optimizing encoders and LLMs to address these issues.

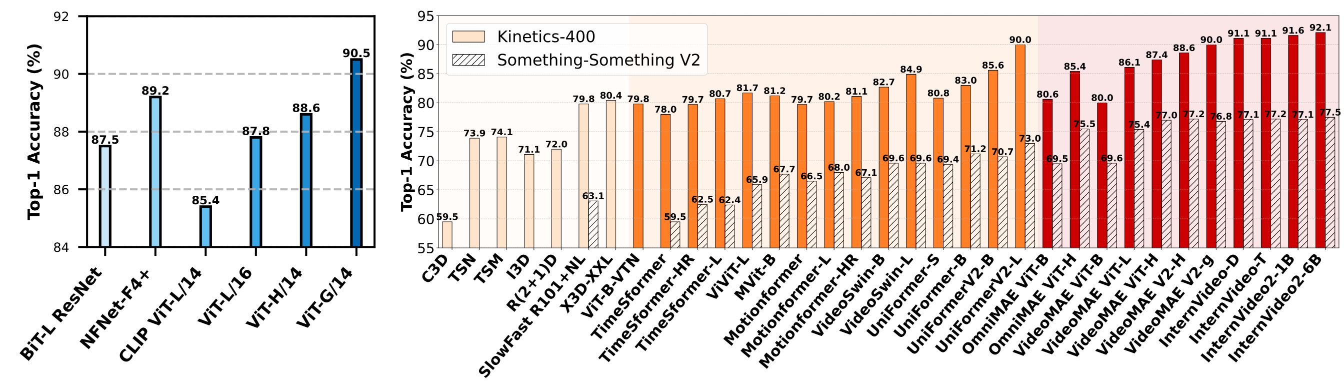
Datasets for video understanding. Datasets are fundamental to advancing video understanding, particularly for assessing the temporal reasoning abilities of LLMs [56, 77, 209]. Current datasets often focus on action recognition, video captioning, or question-answering, capturing spatiotemporal patterns and semantic connections [58, 78, 128]. However, many datasets emphasize short-term motions or provide only surface-level annotations, lacking temporal details such as event order, causality, or duration [82, 88, 137, 175].

While multimodal datasets, combining video with text or audio, offer opportunities for LLMs to align spatiotemporal and semantic reasoning, challenges remain [49, 155]. These include limited diversity in scenarios, imprecise annotations, and the difficulty of representing long-term dependencies [157, 212]. To truly evaluate and enhance LLMs' temporal understanding, future datasets must provide richer, more diverse temporal annotations and robust benchmarks across varied domains [119]. Table 3 provides a comprehensive summary of video datasets used for tasks such as action recognition, video question answering (QA), video captioning, video retrieval, and anomaly detection.

In this work, we emphasize the role of datasets in shaping temporal understanding. We analyze the shortcomings of current video datasets, including their lack of temporal annotations, short-term bias, and limited diversity, and propose directions for improving dataset design.

Temporal reasoning in video AI. Temporal reasoning is critical for tasks such as action recognition, video summarization, and temporal event ordering [21, 26, 29, 44, 84, 121, 125, 156, 160–169, 212]. Classical approaches rely on recurrent neural networks (RNNs) [132], 3D convolutional networks (3D-CNNs) [18, 147], and attention mechanisms [26, 148, 164] to model temporal dependencies. More recently, transformers and temporal tokenization techniques have been explored for long-term video understanding [9, 120]. Despite these advances, the explicit modeling of abstract temporal concepts such as causality, sequence progression, and event duration remains underexplored. LLM-based approaches typically use spatiotemporal embeddings from pretrained encoders but fall short in demonstrating robust temporal reasoning capabilities, especially for complex, real-world video scenarios [157, 212].

While many studies report empirical results [7, 43, 64, 76, 106, 141, 154], we go further by offering actionable insights for advancing LLM-based temporal reasoning. These include joint training approaches, better dataset curation, and innovative multimodal fusion techniques. By addressing these gaps, our work not only advances the understanding of temporal reasoning in LLMs but also provides a roadmap for future research, distinguishing it from prior efforts[118, 142, 210] in this rapidly evolving field.



(a) Image encoders. **(b) Video encoders.**

Figure 2: Performance comparison of visual encoders. (*Left*): Image classification accuracy for various image encoders pretrained and fine-tuned on the ImageNet-1K dataset. (*Right*): Action recognition accuracy for different video encoders pretrained and fine-tuned on the Kinetics-400 and Something-Something V2 datasets.

Task	Dataset	Year	Source	# Videos	Modality	Avg. length (s)	Temporal annotation	Description
Action Recognition	HMDB51[88]	2011	YouTube	6,766	Video	3~4	No	Daily human actions
	UCF101[137]	2012	YouTube	13,320	Video+Audio	7.21	No	Human actions (*e.g.*, sports, daily activities)
	ActivityNet[13]	2015	YouTube	27,801	Video+Text	300~1200	Temporal extent provided	Human-centric activities
	Charades[136]	2016	Crowdsourced	9,848	Video+Text	30.1	Start and end timestamps provided	Household activities
	Kinetics-400[82]	2017	YouTube	306,245	Video	10	No	Human actions (*e.g.*, sports, tasks)
	AVA[60]	2018	Movies	430	Video	Variable	Start and end timestamps provided	Action localization in movie scenes
	Kinetics-600[16]	2018	YouTube	392,622	Video	10	No	Human actions (*e.g.*, sports, tasks)
	Something-Something V2[58]	2018	Crowdsourced	220,847	Video	2~6	Weak	Human-object interactions
	EPIC-KITCHENS[38]	2018	Participant kitchens	432	Video+Text+Audio	~458	Start and end timestamps provided	Large-scale egocentric cooking dataset
	COIN[143]	2019	YouTube	11,827	Video+Text	141.6	Start and end timestamps provided	Comprehensive instructional tasks (*e.g.*, cooking, repair)
	Kinetics-700[17]	2019	YouTube	650,317	Video	10	No	Expanded version of Kinetics-400 and Kinetics-600
	EPIC-KITCHENS-100[39]	2020	Participant kitchens	700	Video+Text+Audio	~514	Start and end timestamps provided	Large-scale egocentric cooking dataset
	Ego4D[59]	2021	Wearable Cameras	3,850 hours	Video+Text+Audio	Variable	Start and end timestamps provided	First-person activities and interactions
	VidSitu[129]	2021	YouTube	29,000	Video+Text	~10	Temporal extent for events provided	Event-centric and causal activity annotations
Video QA	MovieQA[144]	2016	Multiple platforms	408	Video+Text	202.7	Start and end timestamps provided	QA for movie scenes
	TGIF-QA[78]	2016	Tumblr GIFs	56,720	Video+Text	3~5	Action timestamps provided	QA over social media GIFs
	MSVD-QA[187]	2017	YouTube	1,970	Video+Text	27.5	Start and end timestamps provided	QA for actions description
	MSRVTT-QA[187]	2017	YouTube	10,000	Video+Text	15~30	Weak	QA across diverse scenes
	TVQA[89]	2019	TV Shows	21,793	Video+Text	60~90	Start and end timestamps provided	QA over medical dramas, sitcoms, crime shows
	ActivityNet-QA[196]	2019	YouTube	5,800	Video+Text	180	Implicit (derived from ActivityNet)	QA for human-annotated videos
	How2QA[95]	2020	HowTo100M (YouTube)	22,000	Video+Text	60	Temporal extent provided	QA over instructional videos
	YouCookQA[172]	2021	YouCook2 (YouTube)	2,000	Video+Text	316.2	Temporal boundaries provided	Cooking-related instructional QA
	STAR[181]	2021	Human activity datasets	22,000	Video+Text	Variable	Action-level boundaries provided	QA over human-object interactions
	MVBench[94]	2023	Public datasets	3,641	Video+Text	5~35	Start and end timestamps provided	Multi-domain QA (*e.g.*, sports, indoor scenes)
	EgoSchema[116]	2023	Ego4D (Wearable Cameras)	5,063	Video+Text	180	Timestamped narrations provided	Long-form egocentric activities
Video Captioning	YouCook[40]	2013	YouTube	88	Video+Text	180~300	Weak	Cooking instructional videos
	MSR-VTT[189]	2016	YouTube	7,180	Video+Text+Audio	10~30	Weak	General scenarios (*e.g.*, sports, transport)
	ActivityNet Captions[86]	2017	YouTube	20,000	Video+Text	180	Start and end timestamps provided	Dense captions for human-centered activities
	VATEX[175]	2019	YouTube	41,250	Video+Text	~10	Weak	Multilingual descriptions with English-Chinese parallel captions
	HowTo100M[117]	2019	YouTube	1.22M	Video+Text+Audio	390	Subtitle timestamps provided	Instructional video captions
	TVC[90]	2020	TV Shows	108,965	Video+Text	76.2	Start and end timestamps provided	Multimodal video captioning dataset
Video Retrieval	LSMDC[128]	2015	Movies	118,114	Video+Text	4.8	Start and end timestamps provided	Large-scale dataset for movie description tasks
	DiDeMo[69]	2017	Flickr (YFCC100M)	10,464	Video+Text	27.5	Start and end timestamps provided	Moment localization in diverse, unedited personal videos
	FIVR-200K[85]	2019	YouTube	225,960	Video	~120	Start and end timestamps provided	Large-scale incident video retrieval dataset with diverse news events
	TVR[90]	2020	TV Shows	21,793	Video+Text	76.2	Start and end timestamps provided	Video-subtitle multimodal moment retrieval dataset
	TextVR[185]	2023	YouTube	10,500	Video+Text	15	Weak	Cross-modal video retrieval with text reading comprehension
	EgoCVR[73]	2024	Ego4D	2,295	Video+Text	3.9~8.1	Weak	Egocentric dataset for fine-grained composed video retrieval
Anomaly Detection	Subway Entrance[2]	2008	Surveillance cameras	1	Video	4,800	No	Crowd monitoring for unusual event detection at subway entrances
	Subway Exit[2]	2008	Surveillance cameras	1	Video	5,400	No	Crowd monitoring for unusual event detection at subway exits
	CUHK Avenue[111]	2013	Surveillance cameras	15	Video	120	No	Urban avenue scenes with anomalies like running, loitering, *etc.*
	Street Scene[126]	2020	Urban street surveillance	81	Video	582	Spatial and temporal bounding boxes	Urban street anomalies, *e.g.*, jaywalking, loitering, illegal parking, *etc.*
	XD-Violence[183]	2020	Movies and in-the-wild scenes	4,754	Video+Audio	~180	Start and end timestamps provided	Multimodal violence detection covering six violence types
	CUVA[46]	2024	YouTube, Bilibili	1,000	Video+Text	~117	Start and end timestamps provided	Causation-focused anomaly understanding across 42 anomaly types
	MSAD[212]	2024	Online Surveillance	720	Video	~20	Frame-level annotations in test set	Multi-scenario dataset with 14 scenarios
Multimodal video tasks	VIDAL-10M[211]	2023	Multiple platforms	10M	Video+Infrared+Depth+Audio+Text	~20	Weak	Multi-domain retrieval dataset
	Video-MME[49]	2024	YouTube	900	Video+Text+Audio	1017.9	Temporal ranges via certificate length	Comprehensive evaluation benchmark across many domains

Table 3: Comprehensive overview of video datasets across tasks.

3 Analysis and Discussion

Can LLMs understand the concept of time? LLMs exhibit several strengths in processing temporal information. Trained on textual data containing narrations or instructions, they can infer temporal relationships through contextual cues such as "first", "then", and "after". When paired with video encoders, LLMs can process spatiotemporal embeddings, enabling tasks like action recognition and temporal event ordering.

However, LLMs lack direct temporal awareness. Standard models do not inherently model the flow of time unless explicitly trained on sequential video data. Instead, they rely on external encoders to provide temporal structure. Capturing long-term dependencies over extended video sequences is another challenge, as LLMs often operate on tokenized inputs within limited context windows. Additionally, video encoders like 3D CNNs [18, 147] or video transformers [120, 158], which act as the "eyes" of the system, may excel in capturing motion patterns but struggle to generalize abstract temporal concepts like causality or duration.

A significant limitation lies in the representation of visual time. Unlike textual representations, visual cues require explicit modeling of motion and event transitions. This ambiguity underscores the need for improved temporal modeling in both encoders and LLMs.

LLMs applied to videos using pretrained visual encoders. Most existing approaches use pretrained image or video encoders to extract visual or spatiotemporal information, rather than designing entirely new encoders (Table 1 and 2). Pretrained image encoders such as CLIP [122], ResNet [68], and ViT [148] excel in capturing spatial information, while video encoders like I3D [18], SlowFast [48], TimeSformer [9], and Video Swin Transformer [109] are widely used for spatiotemporal feature extraction. These encoders, trained on large-scale datasets such as ImageNet [41, 127] or Kinetics [17, 18, 82], are adept at learning rich feature representations, which can then be fine-tuned for specific tasks [21, 26, 44, 84, 157, 161, 165, 166, 169, 170, 212]. Notably, video encoders incorporate mechanisms to model temporal dependencies, such as optical flow [165], Taylor videos [170], and motion tracking [26, 125], addressing a challenge that LLMs alone cannot handle effectively. Figure 2 illustrates the performance of popular visual encoders on two key tasks: image classification (ImageNet-1K) and video action recognition (Kinetics-400 and Something-Something V2).

The use of pretrained visual encoders offers practical advantages. These encoders are optimized for handling visual and spatiotemporal features, reducing computational overhead and enabling faster convergence. Their modular design also allows them to function as "plug-and-play" components, seamlessly integrating with LLMs for multimodal learning. This modularity ensures that LLM-based frameworks remain adaptable and scalable across diverse applications. However, pretrained encoders are not without limitations. First, encoders trained on general datasets may underperform on domain-specific video tasks [21, 212]. Second, many pretrained encoders prioritize spatial over temporal information, necessitating additional modules, such as temporal transformers, to capture complex temporal dynamics [25, 202]. Lastly, large-scale video datasets required for training such encoders are costly to annotate, limiting their ability to encompass diverse or abstract video content.

Some recent efforts, such as DeepMind's Flamingo [5] and unified architectures like Florence [197] and OmniVL [153], explore custom encoders optimized for multimodal learning. These models aim to balance performance across multiple modalities (image, video, and text) without relying heavily on separate pretrained components [157, 161, 166].

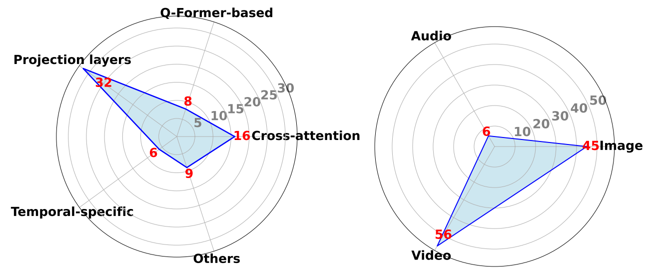
How encoders and LLMs interact? Encoders play a crucial role in preprocessing video frames or sequences to extract visual or spatiotemporal features, which are then transformed into a format compatible with LLMs, often as token embeddings. For example, video transformers process sequences of video frames, while text encoders like CLIP encode textual inputs [6, 191, 201]. These features are projected into a shared representation space to align different modalities, such as visual and textual data.

Fusion mechanisms are pivotal in enabling LLMs to process multimodal inputs. Encoded features are tokenized and treated as input tokens for the LLM. Attention mechanisms, such as cross-modal attention [27, 207], are then used to integrate and relate features from different encoders. Positional embeddings encode spatial and temporal positions of video features, helping the LLM model sequence dependencies. For instance, cross-modal transformers align visual and textual representations, while LLMs refine these embeddings to generate outputs like action labels, video descriptions, or anomaly detection results [47].

Table 1 highlights the interaction/fusion mechanisms used in the latest video-LLMs under the 'Interaction / Fusion mechanism' column. Figure 3a visualizes the distribution of various interaction and fusion techniques, while Figure 3b showcases the modalities used in closely related works from 2024 (notably, text modalities are consistently used across these studies). As shown in these plots, the majority of works use projection layers (*e.g.*, linear projection [94], MLP projection [50, 203, 205], semantic translators [20, 25]) and cross-attention mechanisms (*e.g.*, cross-modal attention [191], gated cross-attention [150]) to facilitate interaction between encoders and LLMs. Temporal-specific mechanisms, such as temporal sampler [202] and scene-level sequential alignment [193], are also used. Emerging video-LLMs are beginning to explore novel fusion mechanisms, including tracklet-centric approach [151], Perceiver resamplers [5, 213], and MQ-Former [152].

However, several challenges arise in this interaction. Encoders must output features in a format compatible with LLMs, requiring careful dimension alignment and embedding space mapping. Additionally, processing video data generates a substantial volume of information, which can strain the LLM's capacity. Capturing long-term dependencies across extended video sequences also remains a challenge, even with support from pretrained encoders.

Bridging the gap between raw video data and temporal reasoning. The interaction between LLMs and encoders helps bridge the gap between raw video data and higher-level temporal reasoning. Pretrained encoders extract spatiotemporal embeddings that encapsulate low-level motion cues, such as velocity [170] and trajectory [120], as well as higher-level temporal patterns like scene transitions and sequence progression. These embeddings provide the foundation for LLMs to interpret complex temporal concepts such as causality, event progression, and anticipation.



(a) Interaction/fusion. **(b) Modalities.**

Figure 3: The distributions of interaction/fusion mechanisms and data modalities in 66 closely related video-LLMs from January 2024 to December 2024. (*Left*): Fusion mechanisms are classified into five categories: Cross-attention (*e.g.*, cross-attention modules, gated cross-attention), Projection layers (*e.g.*, linear projection, MLP projection), Q-Former-based methods (*e.g.*, Q-Former aligns multi-modal embeddings, Trainable Q-Former), Motion/Temporal-Specific mechanisms (*e.g.*, temporal samplers, scene-level sequential alignment), and Other Methods (*e.g.*, Tracklet-centric, Perceiver Resampler, MQ-Former). (*Right*): The distribution of data modalities used in these video-LLMs, with text modalities appearing across all models. Note that a model may use multiple fusion methods and/or data modalities.

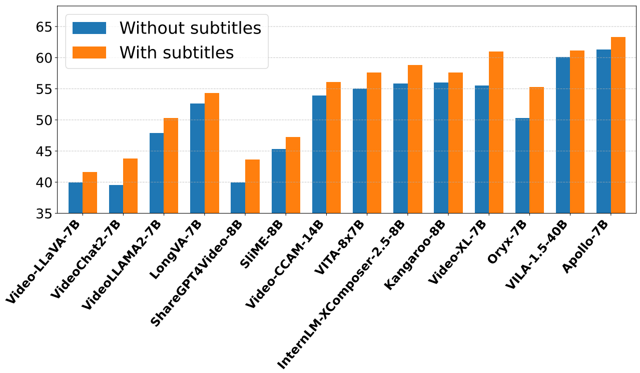


Figure 4: Performance (accuracy) comparison of recent video-LLMs on the Video-MME benchmark.

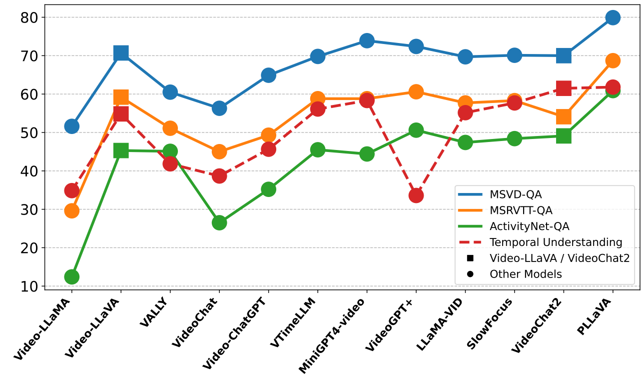


Figure 5: Performance comparison of recent video-LLMs on video QA benchmarks. Models using pretrained video encoders (*e.g.*, Video-LLaVA and VideoChat2) are marked with squares, while models using pretrained image encoders are represented by circles.

LLMs use attention mechanisms to prioritize spatiotemporal features, aligning them with semantic or task-specific contexts. Encoders often supply frame-level features as temporal tokens, enabling the LLM to model dependencies and transitions across frames. However, significant challenges remain in achieving comprehensive temporal understanding. Encoders frequently focus on short-term motion patterns, neglecting long-term dependencies [162, 167, 168]. Similarly, datasets used for encoder pretraining often lack diversity and fail to represent abstract temporal relationships effectively [157]. Table 3 provides a comparison of datasets based on average video length, data source, modalities, and the number of videos.

LLMs also face inherent limitations. Temporal embeddings require careful tokenization to preserve sequence information when input into LLMs (see Table 1). Additionally, LLMs pretrained on static datasets such as text or images may lack the dynamic reasoning capabilities required for video tasks [65]. Combining spatial, temporal, and semantic information without losing critical cues remains a complex challenge [21, 26, 125], further compounded by the computational expense of processing long video sequences while retaining both global and local temporal details.

Video datasets: an enabler or bottleneck? Video datasets are foundational to LLMs on video tasks, yet they often act as a bottleneck [157]. Action recognition datasets, such as Kinetics [16, 82] and Something-Something V2 [58], are effective for analyzing short-term motion patterns but lack the temporal annotations necessary for reasoning about action sequences or causal relationships. Compared to the Something-Something V2 dataset, the Kinetics datasets exhibit a stronger spatial bias, as demonstrated in [55, 57, 120]. Similarly, video QA datasets like TVQA [89] and How2QA [95] align well with LLM architectures but often rely on scripted scenarios that limit generalizability to real-world tasks.

Video captioning datasets like MSR-VTT [189] and ActivityNet Captions [86] enable multimodal learning by fusing video and text embeddings. However, captions are often superficial and fail to probe deeper temporal reasoning. Long-term video understanding datasets, such as Ego4D [59] and Charades [136], focus on extended activities and interactions, offering a richer testing ground for temporal reasoning. Nonetheless, LLMs often struggle with the scale and complexity of such datasets, given their limited context windows.

Multimodal datasets like HowTo100M [117] and COIN [143] align video content with auxiliary modalities, providing opportunities for pretraining in a multimodal setup. However, the inherent noise and lack of temporal annotations in these datasets can hinder performance. To advance temporal understanding in LLMs, datasets must evolve to include richer temporal annotations, long-term dependencies, and diverse real-world scenarios. The latest multimodal datasets include VIDAL-10M [211] and Video-MME [49].

Table 3 summarizes popular datasets used across various video tasks, including action recognition, anomaly detection, video question answering, captioning, and retrieval, as well as some recent multimodal video understanding datasets.

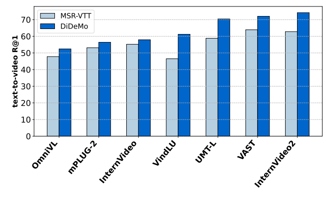
State-of-the-Art video LLMs. Recent advancements in video LLMs have significantly enhanced the processing of spatiotemporal information [22, 94, 119]. Traditional LLMs, focused on textual data, often struggle with temporal dynamics in video [106]. Models like VideoChat2 [94] and SlowFocus [119] are pushing the boundaries by integrating temporal reasoning with video analysis. VideoChat2 enables real-time multimodal dialogue, processing video sequences to answer questions about actions, events, and causal relationships, making it highly effective for interactive applications. SlowFocus improves fine-grained temporal understanding, capturing long-term dependencies and transitions within video, essential for tasks like video summarization and anomaly detection. TimeSformer [9] further refines this by using attention mechanisms to simultaneously model spatial and temporal features, enhancing video understanding in complex scenarios like action recognition.

Additionally, models, *e.g.*, VideoLLM-Online [22] and VideoBERT [139], focus on continuous learning from video streams, allowing for real-time updates and better adaptability to dynamic content. These models are crucial for applications requiring ongoing video analysis, such as surveillance, event detection, and interactive media. Flamingo [5] takes multimodal learning a step further by combining visual, textual, and temporal data, offering a more holistic approach to video processing. ActionFormer [200], on the other hand, specializes in action recognition through long-range temporal dependencies, making it effective for tasks like sports video analysis and human-computer interaction. These advancements reflect a significant leap in video LLM capabilities, making them better equipped for real-world video understanding, interaction, and analysis. Figures 4, 5, and 6 present comparisons of recent popular video-LLMs across multiple video tasks, including multimodal video understanding, video QA, video retrieval, and video captioning. As shown in these figures, no single video-LLM excels across all tasks: (i) a comprehensive evaluation system covering all video tasks is lacking, and (ii) most video-LLMs are designed to address only a subset of these challenges.

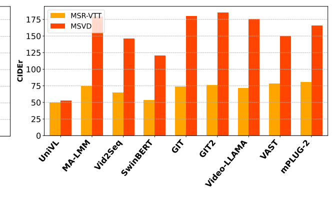
Fair evaluation is needed. Evaluations and comparisons of video-LLMs are often conducted inconsistently (see Figures 4, 5, and 6), which can result in unfair assessments and misleading conclusions [118, 142]. A frequent issue is the comparison of video-LLMs, designed for multimodal reasoning, with traditional video models such as I3D [18], SlowFast [48], or Video Swin Transformer [109], which are tailored for video-specific tasks like action recognition. While traditional models excel at spatiotemporal feature extraction due to their focused design, video-LLMs must simultaneously handle visual and linguistic alignment, which adds inherent complexity to their objectives. Directly comparing video-LLMs against such specialized models is therefore not entirely fair. Instead, video-LLMs should be systematically benchmarked against other video-LLMs or multimodal frameworks to better reflect their relative strengths and limitations in tasks like video action recognition, video captioning, or video QA.

Furthermore, inconsistencies in evaluation practices exacerbate the problem. Different models are often trained and tested on varying datasets, such as Kinetics-400 [82], Ego4D [59], or HowTo100M [117], without standardized protocols for pretraining, finetuning, or testing. This creates biases in results and hampers fair comparisons. For instance, a video-LLM pretrained on massive multimodal datasets might show superior results simply due to larger data availability rather than architectural improvements. To address this, evaluations must adopt consistent and standardized benchmarks, training splits, and metrics. Frameworks that systematically assess multimodal alignment, temporal reasoning, and downstream task performance would help ensure transparency and comparability.

Finally, establishing fair evaluation practices requires prioritizing within-paradigm comparisons. For video action recognition, for example, models like VideoChat2 [94] and VideoLLM-online [22] should be compared against each other rather than with traditional video-only transformers. This approach highlights progress within the multimodal video understanding space and reveals areas for improvement, such as better temporal consistency or more efficient multimodal alignment. By addressing these challenges, fair and systematic evaluation will provide deeper insights into video-LLM capabilities and foster future advancements in the field.



(a) Video retrieval.



(b) Video captioning.

Figure 6: Performance comparison of recent video-LLMs on (a) video retrieval and (b) video captioning benchmarks.

Factors driving superior performance in video LLMs. The superior performance of certain video LLMs can be attributed to their ability to effectively integrate spatial, temporal, and semantic information, often through advanced architectures and training strategies. Models like SlowFocus [119] and VideoChat2 [94] excel by incorporating fine-grained temporal reasoning and long-range dependencies, which are crucial for understanding complex video dynamics such as event progression and causal relationships. The use of hierarchical or multi-level attention mechanisms, seen in models like TimeSformer [9] and VideoLLM-Online [22], enables them to capture both short-term motions and long-term narrative structures, addressing the temporal gaps that limit earlier models.

Additionally, models that use large-scale, multimodal pretraining on diverse video datasets, such as Flamingo, benefit from the cross-domain knowledge transfer between visual, textual, and temporal modalities, leading to a more holistic understanding of video content. These innovations enable the models to generalize better across different video tasks, including action recognition, video captioning, and dynamic scene interpretation.

4 Future Directions

Building on the preceding analysis and discussion, we outline below several promising future research directions for those interested in advancing video LLMs.

Overcoming dataset challenges for LLMs. Datasets remain a critical bottleneck in advancing LLM-based video systems. Addressing their limitations requires both creative solutions and resource investments:

i. Temporal annotations and structure: Enriching datasets with temporal annotations, such as event order, duration, and causal relationships, is essential. Techniques like crowd-sourced annotations, AI-assisted labeling, or synthetic data generation could address the scarcity of such datasets.
ii. Balancing scale and quality: While large-scale datasets like HowTo100M offer broad coverage, they often include noise and irrelevant data. Future efforts should focus on curating high-quality, balanced datasets that prioritize diversity and accuracy without sacrificing scale. Semi-supervised and unsupervised learning approaches could also mitigate the need for large annotated datasets.
iii. Addressing short-term bias: The dominance of short video clips in existing datasets limits models' ability to reason over extended sequences. Introducing datasets with long-term dependencies, such as episodic or procedural videos, would help train models to understand overarching narratives and transitions.
iv. Expanding dataset diversity: Current datasets often focus on narrow domains, such as sports or cooking, which restricts model generalizability [157]. Diverse datasets encompassing a wider range of cultural contexts, real-world scenarios, and tasks are essential to improve performance across applications. Collaborative initiatives between academic and industry stakeholders could accelerate the development of such datasets.

v. Improving multimodal alignment: Misalignment between video frames and associated text or audio annotations introduces inconsistencies that can degrade learning quality. Future work could explore more precise alignment methods [162, 163, 167, 168], such as neural alignment models or reinforcement learning frameworks, to ensure that temporal and semantic signals are accurately correlated during training.

Enhancing temporal understanding. To improve temporal reasoning, joint training of encoders and LLMs is a promising direction. Models co-trained on datasets with temporal reasoning tasks can develop a deeper understanding of complex time-related concepts such as causality, event sequencing, and duration. Architectures like temporal transformers, recurrent neural networks, or hybrid systems that combine hierarchical and sequential processing should be further explored to handle both short-term dynamics and long-term dependencies in video data.

Explicit supervision for abstract temporal concepts through enriched annotations is another critical step [44]. Annotated datasets with detailed temporal labels, covering relationships, transitions, and event causality, can significantly boost the temporal reasoning capacity of these systems. Furthermore, multimodal training on datasets that integrate video, text, audio, and temporal metadata could align spatiotemporal and semantic representations more effectively, enhancing the model's ability to reason about time in real-world scenarios [44, 125, 212].

Multimodal LLMs for holistic temporal reasoning. The development of truly multimodal LLMs capable of holistic temporal reasoning requires a synergistic interplay between spatiotemporal and semantic understanding. Future research should explore adaptive attention mechanisms that dynamically weigh temporal, spatial, and textual information based on context. Advances in memory-efficient architectures and progressive learning strategies could enable LLMs to process longer sequences without losing critical details.

Additionally, fine-tuning LLMs on domain-specific datasets or tasks, such as medical video analysis or video-based social interaction studies [36, 79], could expand their applicability and temporal reasoning depth. Transfer learning approaches [206], where models pretrained on general datasets are fine-tuned for specific temporal reasoning tasks, can also be effective in reducing resource demands.

Other modalities, such as depth videos [156, 159] and motion-specific data like skeletons (useful for analyzing human-related movements) [162–164, 167, 168], optical flow [165], and Taylor videos [170], can significantly enhance the performance of LLM frameworks in video processing tasks when their pretrained models are incorporated. Depth videos capture three-dimensional spatial information, offering a richer understanding of scene geometry. Skeleton data focuses on joint movements, making it particularly effective for applications like action recognition and gesture analysis. Optical flow and Taylor videos excel in capturing frame-to-frame motion changes, providing detailed temporal cues essential for understanding dynamic content.

By integrating these diverse modalities, LLMs can achieve a more comprehensive representation of motion dynamics and spatial structures, broadening their applicability to complex video-based challenges. Moreover, the inclusion of learned video motion prompts [26], which highlight relevant movements within a scene, introduces a novel modality that further refines the system's ability to process and interpret intricate video content.

Advancing visual encoders for multimodal learning. While most current LLM-based video systems use pretrained encoders for their efficiency and robust feature extraction, there is significant potential in designing novel encoders optimized specifically for multimodal learning. These encoders should aim to seamlessly integrate spatial, temporal, and semantic information into a unified framework, reducing the reliance on modular preprocessing steps. Future research could explore adaptive encoder architectures that dynamically adjust to varying video characteristics, such as scene complexity or temporal dynamics. Additionally, encoders tailored for specific domains, like healthcare, education, or autonomous driving, could enhance the accuracy and relevance of multimodal systems in specialized applications.

Ethical and practical considerations. As LLM-based video systems advance, addressing ethical concerns becomes increasingly important [170, 212]. Ensuring fairness and avoiding biases in datasets, particularly regarding cultural and contextual diversity, will be critical. Practical considerations like energy efficiency and the environmental impact of large-scale training should also guide future research. Lightweight models or distillation techniques could balance performance with computational sustainability. Addressing these directions, researchers can unlock the full potential of LLMs in video-based applications, driving advancements in temporal reasoning, multimodal understanding, and real-world usability.

5 Conclusion

This work critically examines the temporal reasoning capabilities of large language models (LLMs) in video processing, identifying significant limitations in both models and datasets. While LLMs paired with pretrained visual encoders have achieved success in tasks such as action recognition, anomaly detection, and video summarization, they fall short in understanding long-term temporal dependencies. This stems from the encoders' focus on short-term patterns, fragmented temporal cues, and challenges in aligning spatial, temporal, and semantic information. Additionally, existing datasets lack explicit temporal annotations, often focus on short clips over long sequences, and struggle with diversity and multimodal alignment, further hindering progress.

To unlock the full potential of LLMs in video processing, future research must address these gaps. This includes designing integrated frameworks to jointly train encoders and LLMs on temporal reasoning, enriching datasets with detailed annotations for long-term dependencies, and creating innovative architectures that fuse spatiotemporal and semantic information. By addressing these challenges, we can pave the way for systems that not only excel in video analysis but also advance broader applications requiring robust temporal comprehension.

Acknowledgments

Xi Ding, a Research Assistant with the Temporal Intelligence and Motion Extraction (TIME) Lab at ANU, contributed to this work. This research was conducted under the supervision of Lei Wang.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219

[2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. 2008. Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 3 (2008), 555–560. https://doi.org/10.1109/TPAMI.2007.70825

[3] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652 [cs.CL] https://arxiv.org/abs/2403.04652

[4] Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[6] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413* (2024).

[7] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. 2023. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2503–2516.

[8] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

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

[10] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. 2021. High-performance large-scale image recognition without normalization. In *International conference on machine learning*. PMLR, 1059–1071.

[11] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics* 9 (2021), 978–994.

[12] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

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

[14] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, and Xiangyu Zhang. 2022. Reversible column networks. *arXiv preprint arXiv:2212.11696* (2022).

[15] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. InternLM2 Technical Report. arXiv:2403.17297 [cs.CL]

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

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

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

[19] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3558–3568.

[20] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. 2023. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292* (2023).

[21] Huilin Chen, Lei Wang, Yifan Chen, Tom Gedeon, and Piotr Koniusz. 2024. When Spatial meets Temporal in Action Recognition. *arXiv preprint arXiv:2411.15284* (2024).

[22] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 2024. VideoLLM-online: Online Video Large Language Model for Streaming Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18407–18418.

[23] Jin Chen, Xinxiao Wu, Yao Hu, and Jiebo Luo. 2021. Spatial-temporal causal inference for partial image-to-video adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1027–1035.

[24] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325* (2024).

[25] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. 2024. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. *arXiv preprint arXiv:2405.20340* (2024).

[26] Qixiang Chen, Lei Wang, Piotr Koniusz, and Tom Gedeon. [n. d.]. Motion meets attention: Video motion prompts. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

[27] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems* 36 (2023), 72842–72866.

[28] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058* (2022).

[29] Wenshuo Chen, Hongru Xiao, Erhang Zhang, Lijie Hu, Lei Wang, Mengyuan Liu, and Chen Chen. 2024. SATO: Stable Text-to-Motion Framework. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6989–6997.

[30] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794* (2022).

[31] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271* (2024).

[32] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821* (2024).

[33] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[34] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476* (2024).

[35] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)* 2, 3 (2023), 6.

[36] Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024. Med42–Evaluating Fine-Tuning Strategies for Medical LLMs: Full-Parameter vs. Parameter-Efficient Approaches. *arXiv preprint arXiv:2404.14779* (2024).

[37] StableLM contributors. 2023. StableLM: Stability AI language models. https://github.com/stability-AI/stableLM

[38] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. *ArXiv* abs/1804.02748 (2018). https://api.semanticscholar.org/CorpusID:4710439

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

[40] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2634–2641.

[41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/V1/N19-1423

[43] Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics* 10 (03 2022), 257–273. https://doi.org/10.1162/tacl_a_00459 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00459/2004543/tacl_a_00459.pdf

[44] Dexuan Ding, Lei Wang, Liyun Zhu, Tom Gedeon, and Piotr Koniusz. 2024. Lego: Learnable expansion of graph operators for multi-modal feature fusion. *arXiv preprint arXiv:2410.01506* (2024).

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

[46] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. 2024. Uncovering What, Why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly. arXiv:2405.00181 [cs.CV] https://arxiv.org/abs/2405.00181

[47] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. 2024. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023* (2024).

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

[49] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075* (2024).

[50] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211* (2024).

[51] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. 2024. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261* (2024).

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

[53] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214* (2024).

[54] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

[55] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10406–10417.

[56] Rohit Girdhar and Deva Ramanan. 2019. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. *arXiv preprint arXiv:1910.04744* (2019).

[57] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. 2022. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16102–16112.

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

[59] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.

[60] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6047–6056.

[61] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems* 35 (2022), 26418–26431.

[62] Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Xi Chen, and Bo Zhao. 2024. VTG-LLM: Integrating Timestamp Knowledge into Video LLMs for Enhanced Video Temporal Grounding. *arXiv preprint arXiv:2405.13382* (2024).

[63] Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. 2024. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643* (2024).

[64] Wes Gurnee and Max Tegmark. 2024. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=jE8xbmvFin

[65] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad ii: The sequel-who, when, and what in movie audio description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13645–13655.

[66] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD III: The Prequel-Back to the Pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18164–18174.

[67] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. 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*. 13504–13514.

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

[69] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing Moments in Video with Natural Language. arXiv:1708.01641 [cs.CV] https://arxiv.org/abs/1708.01641

[70] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).

[71] Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353* (2024).

[72] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition. 14271–14280.

[73] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. 2024. EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval. *arXiv preprint arXiv:2407.16658* (2024).

[74] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video ReCap: Recursive Captioning of Hour-Long Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18198–18208.

[75] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017* (2022).

[76] Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do Language Models Have a Common Sense regarding Time? Revisiting Temporal Commonsense Reasoning in the Era of Large Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://openreview.net/forum?id=akJUrevmwI

[77] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2019. Video question answering with spatio-temporal reasoning. *International Journal of Computer Vision* 127 (2019), 1385–1412.

[78] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2758–2766.

[79] Cheonsu Jeong. 2024. Fine-tuning and utilization methods of domain-specific llms. *arXiv preprint arXiv:2401.02981* (2024).

[80] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[81] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

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

[83] Byoungjip Kim, Dasol Hwang, Sungjun Cho, Youngsoo Jang, Honglak Lee, and Moontae Lee. 2024. Show Think and Tell: Thought-Augmented Fine-Tuning of Large Language Models for Video Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1808–1817.

[84] Piotr Koniusz, Lei Wang, and Anoop Cherian. 2021. Tensor representations for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 2 (2021), 648–665.

[85] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. 2019. FIVR: Fine-grained incident video retrieval. *IEEE Transactions on Multimedia* 21, 10 (2019), 2638–2652.

[86] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.

[87] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[88] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*. IEEE, 2556–2563.

[89] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696* (2018).

[90] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 447–463.

[91] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:204960716

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

[93] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).

[94] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.

[95] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200* (2020).

[96] Yanwei Li, Chengyao Wang, and Jiaya Jia. 2025. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*. Springer, 323–340.

[97] Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. 2024. VideoINSTA: Zero-shot Long Video Understanding via Informative Spatial-Temporal Reasoning with LLMs. *arXiv preprint arXiv:2409.20365* (2024).

[98] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).

[99] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26689–26699.

[100] Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. 2022. A causal inference look at unsupervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1620–1629.

[101] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, Xinchao Wang, and Thomas S. Huang. 2018. Learning Temporal Dynamics for Video Super-Resolution: A Deep Learning Approach. *IEEE Transactions on Image Processing* 27, 7 (2018), 3432–3445. https://doi.org/10.1109/TIP.2018.2820807

[102] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 [cs.CV] https://arxiv.org/abs/2310.03744

[103] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV] https://arxiv.org/abs/2304.08485

[104] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542* (2024).

[105] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2025. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*. Springer, 1–18.

[106] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476* (2024).

[107] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3042–3051.

[108] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2024. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961* (2024).

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

[110] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. 2024. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256* (2024).

[111] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal Event Detection at 150 FPS in MATLAB. In *2013 IEEE International Conference on Computer Vision*. 2720–2727. https://doi.org/10.1109/ICCV.2013.338

[112] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207* (2023).

[113] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093* (2023).

[114] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding. *arXiv preprint arXiv:2406.09418* (2024).

[115] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).

[116] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems* 36 (2023), 46212–46244.

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

[118] Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024. Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives. *arXiv preprint arXiv:2406.05615* (2024).

[119] Ming Nie, Dan Ding, Chunwei Wang, Yuanfan Guo, Jianhua Han, Hang Xu, and Li Zhang. 2024. SlowFocus: Enhancing Fine-grained Temporal Understanding in Video LLM. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=FOkKndty5B

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

[121] Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, Saeed Anwar, and Tom Gedeon. 2022. Fusing higher-order features in graph neural networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems* 35, 4 (2022), 4783–4797.

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

[123] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[124] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[125] Arjun Raj, Lei Wang, and Tom Gedeon. 2024. TrackNetV4: Enhancing Fast Sports Object Tracking with Motion Attention Maps. *arXiv preprint arXiv:2409.14543* (2024).

[126] Bharathkumar Ramachandra and Michael Jones. 2020. Street Scene: A new dataset and evaluation protocol for video anomaly detection. arXiv:1902.05872 [cs.CV] https://arxiv.org/abs/1902.05872

[127] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. ImageNet-21K Pretraining for the Masses. arXiv:2104.10972 [cs.CV]

[128] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision* 123 (2017), 94–120.

[129] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5589–5600.

[130] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[131] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[132] Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404 (2020), 132306.

[133] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. 2024. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998* (2024).

[134] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720* (2023).

[135] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. 2024. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485* (2024).

[136] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 510–526.

[137] K Soomro. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[138] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2443–2449. https://doi.org/10.1145/3404835.3463257

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

[140] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389* (2023).

[141] Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. 2024. Are Language Models Actually Useful for Time Series Forecasting?. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=DV15UbHCY1

[142] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432* (2023).

[143] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1207–1216.

[144] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4631–4640.

[145] Together.xyz. 2023. Releasing 3b and 7b redpajama incite family of models including base, instruction-tuned and chat models. https://www.together.xyz/blog/redpajama-models-v1

[146] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[147] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[148] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[149] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. *Advances in neural information processing systems* 29 (2016).

[150] Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. 2024. COSMO: COntrastive Streamlined MultimOdal Model with Interleaved Pre-Training. *arXiv preprint arXiv:2401.00849* (2024).

[151] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407* (2023).

[152] Junke Wang, Dongdong Chen, Chong Luo, Bo He, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2024. Omnivid: A generative framework for universal video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18209–18220.

[153] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. 2022. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems* 35 (2022), 5696–5710.

[154] Jiexin Wang, Adam Jatowt, and Yi Cai. 2024. Towards Effective Time-Aware Language Representation: Exploring Enhanced Temporal Understanding in Language Models. *arXiv preprint arXiv:2406.01863* (2024).

[155] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, et al. 2024. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319* (2024).

[156] Lei Wang. 2021. Analysis and evaluation of Kinect-based action recognition algorithms. *arXiv preprint arXiv:2112.08626* (2021).

[157] Lei Wang. 2023. *Robust human action modelling*. Ph. D. Dissertation. The Australian National University (Australia).

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

[159] Lei Wang, Du Q Huynh, and Piotr Koniusz. 2019. A comparative review of recent kinect-based action recognition algorithms. *IEEE Transactions on Image Processing* 29 (2019), 15–28.

[160] Lei Wang, Du Q Huynh, and Moussa Reda Mansour. 2019. Loss switching fusion with similarity search for video classification. In *2019 IEEE international conference on image processing (ICIP)*. IEEE, 974–978.

[161] Lei Wang and Piotr Koniusz. 2021. Self-supervising action recognition by statistical moment and subspace descriptors. In *Proceedings of the 29th ACM international conference on multimedia*. 4324–4333.

[162] Lei Wang and Piotr Koniusz. 2022. Temporal-viewpoint transportation plan for skeletal few-shot action recognition. In *Proceedings of the Asian Conference on Computer Vision*. 4176–4193.

[163] Lei Wang and Piotr Koniusz. 2022. Uncertainty-dtw for time series and sequences. In *European Conference on Computer Vision*. Springer, 176–195.

[164] Lei Wang and Piotr Koniusz. 2023. 3mformer: Multi-order multi-mode transformer for skeletal action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5620–5631.

[165] Lei Wang and Piotr Koniusz. 2024. Flow dynamics correction for action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3795–3799.

[166] Lei Wang, Piotr Koniusz, and Du Q Huynh. 2019. Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8698–8708.

[167] Lei Wang, Jun Liu, and Piotr Koniusz. 2021. 3D Skeleton-based Few-shot Action Recognition with JEANIE is not so Naïve. *arXiv preprint arXiv:2112.12668* (2021).

[168] Lei Wang, Jun Liu, Liang Zheng, Tom Gedeon, and Piotr Koniusz. 2024. Meet JEANIE: a Similarity Measure for 3D Skeleton Sequences via Temporal-Viewpoint Alignment. *International Journal of Computer Vision* (2024), 1–32.

[169] Lei Wang, Ke Sun, and Piotr Koniusz. 2024. High-order tensor pooling with attention for action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3885–3889.

[170] Lei Wang, Xiuyuan Yuan, Tom Gedeon, and Liang Zheng. [n. d.]. Taylor Videos for Action Recognition. In *Forty-first International Conference on Machine Learning*.

[171] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[172] Shaojie Wang, Wentian Zhao, Ziyi Kou, Jing Shi, and Chenliang Xu. 2021. How to make a blt sandwich? learning vqa towards understanding web instructional videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1130–1139.

[173] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems* 36 (2024).

[174] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024. Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. *arXiv preprint arXiv:2411.10442* (2024).

[175] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4581–4591.

[176] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. arXiv:2307.06942 [cs.CV] https://arxiv.org/abs/2307.06942

[177] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. 2024. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv e-prints* (2024), arXiv–2403.

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

[179] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. 2024. Loong: Generating Minute-level Long Videos with Autoregressive Language Models. *arXiv preprint arXiv:2410.02757* (2024).

[180] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. 2024. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3907–3916.

[181] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711* (2024).

[182] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. arXiv:2303.04671 [cs.CV] https://arxiv.org/abs/2303.04671

[183] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. arXiv:2007.04687 [cs.CV] https://arxiv.org/abs/2007.04687

[184] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519* (2023).

[185] Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. 2023. A Large Cross-Modal Video Retrieval Dataset with Reading Comprehension. arXiv:2305.03347 [cs.CV] https://arxiv.org/abs/2305.03347

[186] Bo Xu and Mu-ming Poo. 2023. Large language models and brain-inspired general intelligence. *National Science Review* 10, 10 (2023), nwad267.

[187] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. [n. d.]. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *ACM Multimedia*.

[188] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*. PMLR, 38728–38748.

[189] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[190] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994* (2024).

[191] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. 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*. 10714–10726.

[192] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).

[193] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. 2024. Vript: A Video Is Worth Thousands of Words. *arXiv preprint arXiv:2406.06040* (2024).

[194] Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442* (2024).

[195] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. 2024. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. arXiv:2310.05737 [cs.CV] https://arxiv.org/abs/2310.05737

[196] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9127–9134.

[197] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021).

[198] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing Large Language Models for Training-free Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18527–18536.

[199] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11975–11986.

[200] Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*. Springer, 492–510.

[201] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:259075356

[202] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2024. Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM. *arXiv preprint arXiv:2406.12235* (2024).

[203] Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Task Me Anything. *arXiv preprint arXiv:2406.11775* (2024).

[204] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. arXiv:2301.06052 [cs.CV] https://arxiv.org/abs/2301.06052

[205] Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024. Analyzing Temporal Complex Events with Large Language Models? A Benchmark towards Temporal, Long Context Understanding. *arXiv preprint arXiv:2406.02472* (2024).

[206] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19632–19642.

[207] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6586–6597.

[208] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685

[209] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*. 803–818.

[210] Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. 2024. A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038* (2024).

[211] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852* (2023).

[212] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. 2024. Advancing Video Anomaly Detection: A Concise Review and a New Dataset. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[213] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. 2024. Apollo: An Exploration of Video Understanding in Large Multimodal Models. arXiv:2412.10360 [cs.CV] https://arxiv.org/abs/2412.10360