

Video-LMM Post-Training: A Deep Dive into Video Reasoning with Large Multimodal Models

Yolo Yunlong Tang¹, Jing Bi¹, Pinxin Liu¹, Zhenyu Pan², Zhangyun Tan¹, Qianxiang Shen¹, Jiani Liu¹, Hang Hua¹, Junjia Guo¹, Yunzhong Xiao³, Chao Huang¹, Zhiyuan Wang⁴, Susan Liang¹, Xinyi Liu¹, Yizhi Song⁵, Junhua Huang⁶, Jia-Xing Zhong⁷, Bozheng Li⁸, Daiqing Qi⁹, Ziyun Zeng¹, Ali Vosoughi¹, Luchuan Song¹, Zeliang Zhang¹, Daiki Shimada¹⁰, Han Liu², Jiebo Luo¹, Chenliang Xu¹

¹ University of Rochester ² Northwestern University ³ CMU ⁴ UCSB ⁵ Purdue University ⁶ UCLA

⁷ University of Oxford ⁸ Brown University ⁹ University of Virginia ¹⁰ Sony Group Corporation

✉ yunlong.tang@rochester.edu [🔗 yunlong10/Awesome-Video-LMM-Post-Training](https://github.com/yunlong10/Awesome-Video-LMM-Post-Training)

Abstract | Video understanding represents the most challenging frontier in computer vision, requiring models to reason about complex spatiotemporal relationships, long-term dependencies, and multimodal evidence. The recent emergence of Video-Large Multimodal Models (Video-LMMs), which integrate visual encoders with powerful decoder-based language models, has demonstrated remarkable capabilities in video understanding tasks. However, the critical phase that transforms these models from basic perception systems into sophisticated reasoning engines—post-training—remains fragmented across the literature. This survey provides the first comprehensive examination of post-training methodologies for Video-LMMs, encompassing three fundamental pillars: supervised fine-tuning (SFT) with chain-of-thought, reinforcement learning (RL) from verifiable objectives, and test-time scaling (TTS) through enhanced inference computation. We present a structured taxonomy that clarifies the roles, interconnections, and video-specific adaptations of these techniques, addressing unique challenges such as temporal localization, spatiotemporal grounding, long video efficiency, and multimodal evidence integration. Through systematic analysis of representative methods, we synthesize key design principles, insights, and evaluation protocols while identifying critical open challenges in reward design, scalability, and cost-performance optimization. We further curate essential benchmarks, datasets, and metrics to facilitate rigorous assessment of post-training effectiveness. This survey aims to provide researchers and practitioners with a unified framework for advancing Video-LMM capabilities. Additional resources and updates are maintained at: <https://github.com/yunlong10/Awesome-Video-LMM-Post-Training>.

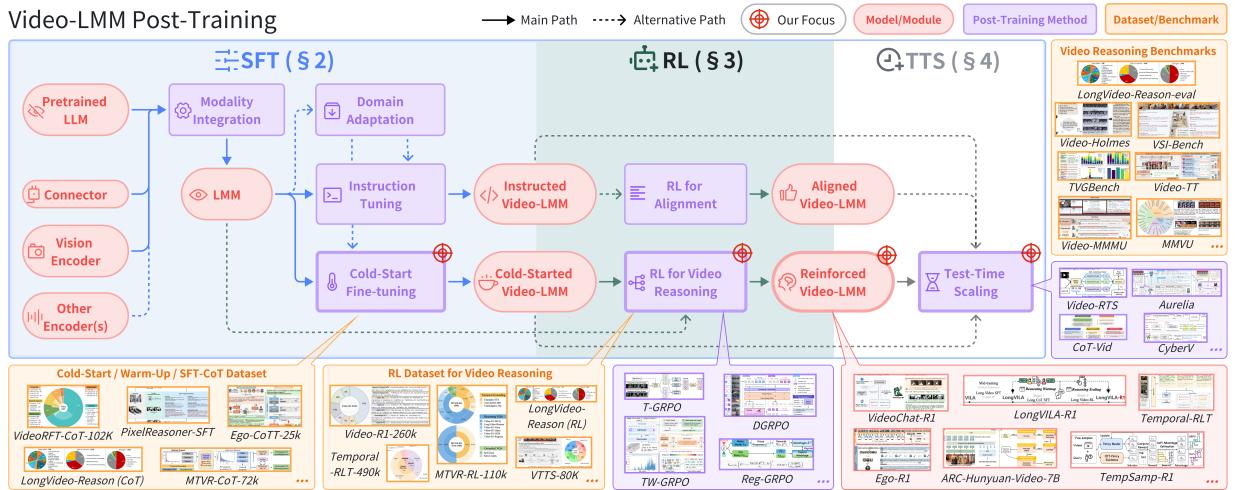


Figure 1 | Overview of Video-LMM post-training and the scope of this survey.

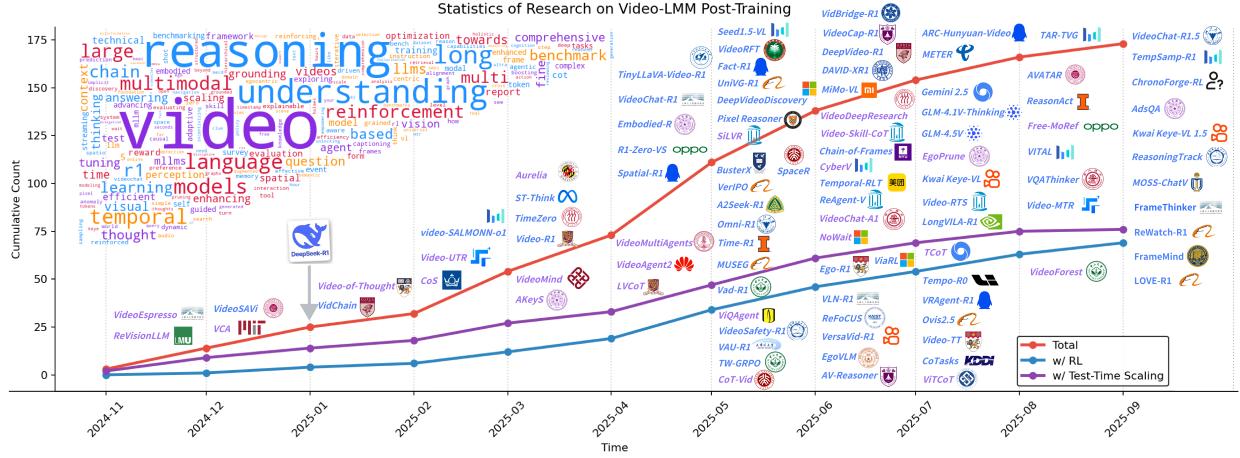


Figure 2 | Research trends in Video-LMM post-training (November 2024 - September 2025). The word cloud is based on the titles of the papers.

1. Introduction

One whale falls, ten thousand beings grow.

— A modern saying, inspired by *The Practice of the Wild* [1]

In recent years, Large Multimodal Models (LMMs) [2–6] have rapidly evolved from simple question-answering toward general problem-solving with interpretable long chain-of-thought (CoT) reasoning [7]. Video understanding, as one of the most comprehensive and challenging directions in computer vision, simultaneously involves complex spatiotemporal relationships, event causality, and long-term memory mechanisms, naturally demanding powerful language reasoning and task interface capabilities [8]. Consequently, Video-LMMs featuring decoder-centric architectures have become the dominant paradigm [8, 9]. These systems leverage strong LLMs as reasoning engines, employ video encoders to extract visual representations, align visual features to the LLM token embedding space through projection modules, and enable instruction understanding and answer generation, demonstrating superior initialization performance and generalization [8].

Video-language modeling has undergone three paradigm shifts: (1) the CNN+RNN era focused on temporal feature aggregation through recurrent architectures [10]; (2) Transformer-based video models, especially BERT-style/encoder-only joint representations, emphasized cross-modal alignment and retrieval through bidirectional encoding [11, 12]; (3) the current video encoder + decoder-based LLM architecture prioritizes the generality and composability of the language component while maximally reusing the knowledge and reasoning capabilities of pretrained LLMs [8, 9, 13]. The key advantage lies in internet-scale self-supervised learning in the language domain, where next-token prediction enables knowledge, reasoning, and interface capabilities to emerge at scale under a unified objective. In contrast, the visual domain lacks an equivalent self-supervised learning method for efficiently processing internet-scale native video data. Although native multimodal approaches that jointly model vision and language end-to-end are being explored [14, 15], they have yet to surpass the divide-and-conquer strategy in computational efficiency and engineering reusability.

Within this framework, **post-training** is the critical phase determining whether Video-LMMs progress from basic perception to sophisticated reasoning. As illustrated in Figure 1, post-training encompasses three major components: (i) Supervised Fine-Tuning (SFT) incorporates CoT and

reasoning style distillation to bootstrap reasoning formats and establish task-following behaviors [16–18]. (ii) Reinforcement Learning (RL) has evolved from RLHF, PPO, and DPO to R1-style/GRPO [19] approaches that eliminate the need for preference data and explicit reward models, enabling enhanced reasoning and self-correction through verifiable objectives and systematic exploration [20–23]. (iii) Test-Time Scaling (TTS) leverages increased inference computation for higher reliability through reasoning sample augmentation, voting mechanisms, self-consistency checks, external verifiers, and multi-path search [24–26]. This progression maintains close alignment with LLM community developments, offering transferability in theoretical principles and engineering practices.

Adapting these paradigms to video presents distinctive challenges differing substantially from static image-text scenarios. Temporal localization requires models to provide not only correct answers but also temporally precise responses anchored to specific segments [27, 28]. Spatiotemporal grounding demands consistency in tracking objects, parts, and actions across spatial and temporal dimensions [22, 29]. Long video understanding necessitates sophisticated sampling strategies, adaptive routing, hierarchical viewing protocols, and effective caching [30, 31]. Multimodal evidence integration requires coordinated reasoning over video frames, textual captions, audio transcripts, and external knowledge [32–34]. These characteristics have catalyzed video-specific post-training strategies: incorporating verifiable temporal and spatial rewards (tIoU, region consistency metrics) in RL frameworks; designing TTS methods that guide models to autonomously select informative frames and perform staged viewing with multi-round reflection and self-correction; and unifying diverse tasks (question answering, temporal localization, spatiotemporal grounding) within coherent alignment and optimization frameworks, establishing hierarchical pipelines for watching, thinking, locating, and answering [24, 27, 28].

Recent studies have successfully integrated GRPO/R1-style RL with extended reasoning TTS into video understanding, as illustrated in Figure 2. Some emphasize verifiable reward design for temporal reasoning and localization [27, 28], others extend to joint spatiotemporal grounding [22], while others focus on long video scaling with efficient training and inference [30, 35], and interactive viewing paradigms enabling thinking with video through evidence accumulation across iterations [24, 36, 37]. This research wave has validated the feasibility of transferring LLM post-training paradigms to video understanding and revealed common challenges in data construction, reward robustness, evaluation protocol standardization, and cost-performance optimization, underscoring the need for a comprehensive survey examining Video-LMM reasoning methods from a post-training perspective.

In this survey, we focus on post-training for Video-LMMs, providing systematic coverage of key techniques across SFT, RL, and TTS, along with their specialized adaptations for video scenarios. We synthesize design principles and engineering insights from representative methods and discuss open challenges and future directions under unified evaluation and reporting standards.

In short, the key contributions of this survey are as follows:

Contributions

- A comprehensive review of post-training methodologies for Video-LMMs, covering supervised fine-tuning, reinforcement learning, and test-time scaling as integral components of model optimization.
- A structured taxonomy of Video-LMM post-training techniques, clarifying their functional roles and interconnections, with insights into open challenges and future directions.
- Practical guidance introducing essential benchmarks, datasets, and evaluation metrics for assessing Video-LMM post-training effectiveness.

Related Surveys. Several surveys have reviewed video understanding with large language models [8, 9, 38–40], multimodal chain-of-thought reasoning [7], and reinforcement learning in LMMs [20]. We also note the recent survey on reinforcement learning for large reasoning models [41], which provides broader context on RL-driven reasoning complementary to video post-training. While these works provide valuable perspectives on video-language modeling and reasoning techniques, our survey distinctly focuses on systematic organization and analysis of post-training methodologies specifically tailored for Video-LMMs, offering a unified treatment of SFT, RL, and TTS approaches.

Survey Structure. Section 2 examines SFT for effective Video-LMM fine-tuning, especially CoT-SFT. Section 3 reviews LLM-based RL foundations before systematically analyzing RL algorithms, especially R1-style methods for video reasoning, including model configurations, data preparation, optimization strategies, and policy/reward design. Section 4 investigates video-specific TTS methods, emphasizing adaptive viewing mechanisms, multi-path reasoning strategies, and verification architectures. Section 5 surveys datasets, benchmarks, and evaluation metrics. Section 6 discusses future directions. Additional resources and updates are maintained at: <https://github.com/yunlong10/Awesome-Video-LMM-Post-Training>.

2. Supervised Fine-Tuning for Video Reasoning

Supervised fine-tuning (SFT) serves as a pivotal stage that not only refines multimodal alignment, enhances instruction-following capability, and instills structured reasoning behaviors in Video-LMMs but also bridges large-scale pretraining and reinforcement learning (RL), laying the foundation for stable and generalizable video reasoning.

Takeaways

- Fixed-format CoT supervision enables imitation of reasoning patterns but provides limited flexibility for self-exploration and error correction compared to RL approaches, necessitating the transition to RL for learning abstract objectives and generalizing to complex, unseen scenarios.
- CoT-SFT has evolved from a standalone training paradigm to a critical cold-start phase for RL, providing structured reasoning formats (<think>, <answer>) and stable initialization that prevents instability in subsequent RL-driven policy optimization.

2.1. Basic SFT for Video-LMMs

Researchers have discovered large-scale pretraining methods that enable LLMs to effectively consume internet-scale unlabeled text corpora through next token prediction, trained with maximum likelihood estimation (MLE) to obtain powerful LLM base models. These base models are then further refined through SFT using high-quality annotated data in smaller quantities. Early SFT for text-only LLMs primarily served two purposes: enhancing the model’s instruction-following capability and performing domain adaptation to transform general-purpose LLMs into domain-specific experts. For obtaining an LMM, subsequent SFT can either build upon the LLM base model or start from an instruction-tuned LLM for further refinement.

Modality Integration. The transition from LLM to LMM typically begins with a Modality Integration stage, which endows the LLM with the ability to understand information from other modalities, particularly visual information. This stage usually employs large-scale image-text pairs for image captioning tasks, sometimes incorporating video-text pairs as well. A connector links the vision encoder to the LLM, and supervised fine-tuning is applied to update either the connector parameters

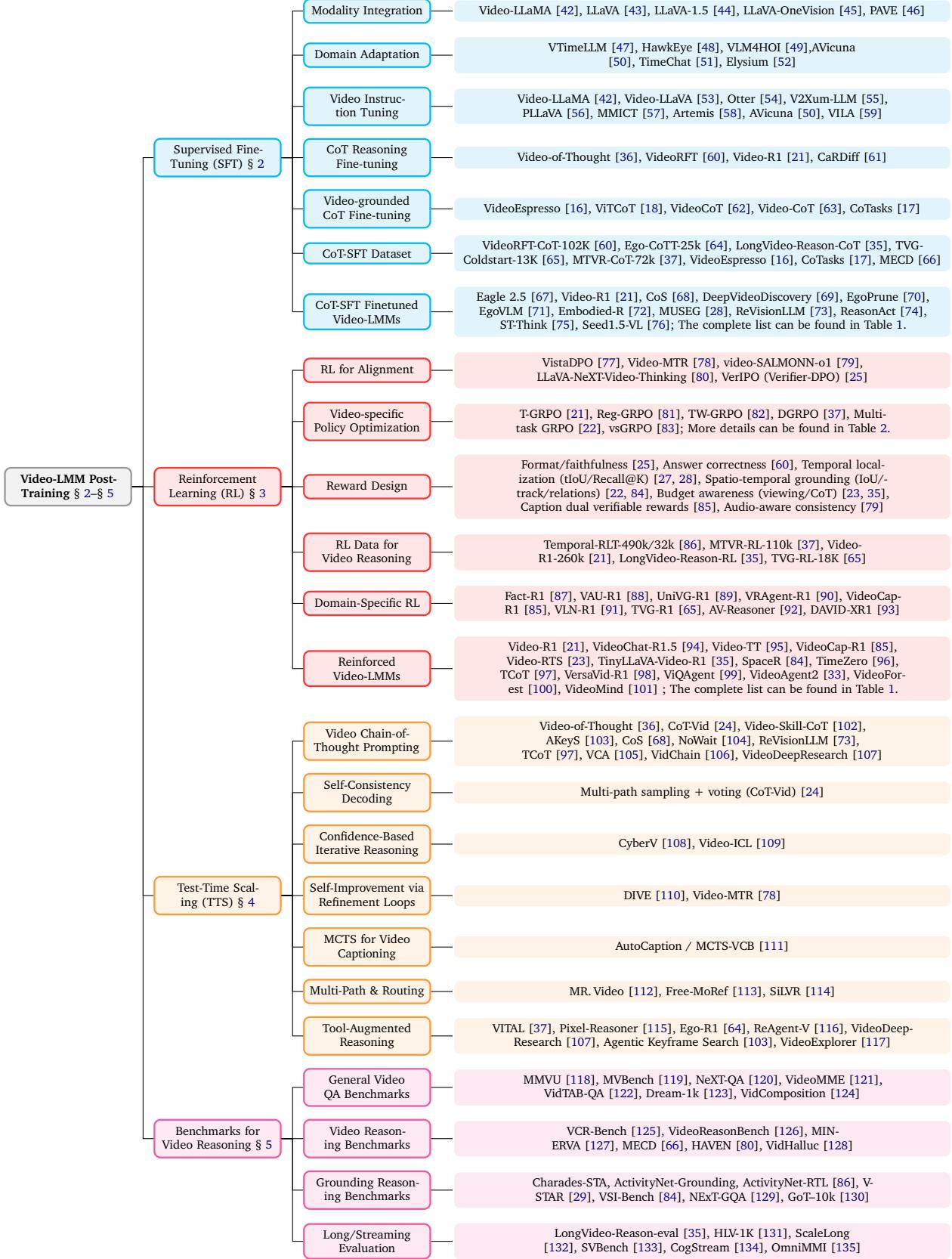


Figure 3 | Taxonomy of Video-LMM post-training.

alone or both the connector and LLM parameters jointly. The connector is typically a linear layer or MLP that maintains input-output token correspondence, though alternatives like Q-Former [136] use resamplers to map inputs to a fixed number of tokens. In practice, the former approach generally outperforms the latter [43]. Additionally, some methods directly feed vision features to the LLM, potentially passing representations from different ViT layers to corresponding LLM layers. Regardless of the specific approach, the key objective of modality integration is to effectively project visual representations from the vision encoder into the LLM’s embedding space, enabling the LMM to directly interpret visual information. Beyond vision, other modalities such as audio, speech, and optical flow can be aligned with LMMs using similar operations [42].

Domain Adaptation. Domain adaptation in Video-LMMs can be understood in multiple ways. The most fundamental interpretation applies when an LMM has only performed modality integration on image-text data without extending to video: an additional domain adaptation step uses video-text pairs to fine-tune the LMM for video captioning, thereby expanding the LMM’s capabilities to video understanding. A second interpretation involves a Video-LMM that initially handles only general video understanding being fine-tuned with domain-specific data to inject domain knowledge, enabling it to process specialized content such as medical videos, anomaly detection videos, or AI-generated video detection. A third interpretation involves endowing Video-LMMs with specific capabilities, such as temporal localization abilities. For instance, VTimeLLM [47], TimeChat [137], and AVicuna [50] employ boundary alignment to align events occurring in videos with their start and end times, enabling LMMs to predict when events occur in videos. Elysium [52] extends this capability to the spatiotemporal domain. Research indicates that domain adaptation may compromise the instruction-following ability inherited from the LLM, typically necessitating further SFT to restore this capability.

Video Instruction Tuning. Video Instruction tuning enhances the instruction-following capability of Video-LMMs [56, 138–144]. The training data takes the form of instruction-response pairs, and after fine-tuning, the model is expected to respond as accurately as possible to any given instruction [145]. For example, when asked to provide a video-to-text summarization of a video, the model generates a description; when asked for video-to-video summarization, the model outputs the indices of key frames [55]. Visual instruction tuning originated with LLaVA [43] and typically follows the Modality Integration stage, though some work has shown that mixing modality integration data with instruction tuning data in a unified format yields better results [44]. Video-LLaMA extended instruction tuning to video and audio, validating the feasibility of video instruction tuning [42]. Since then, instruction tuning has been widely applied to video understanding [30, 146–159].

These fine-tuning approaches all employ auto-regressive language modeling loss as the objective function. While full fine-tuning of the LLM is possible, it can be computationally and memory-intensive, leading to frequent adoption of parameter-efficient fine-tuning (PEFT) techniques. For example, some approaches only update LoRA [160] and connector parameters, while others attempt to fine-tune the vision encoder. Input prompts typically include video placeholders that are replaced with corresponding video tokens before being fed into the LLM.

2.2. From Video Instruction Tuning to Chain-of-Thought Fine-tuning (CoT-SFT)

CoT Reasoning Fine-tuning. Chain-of-thought (CoT) reasoning emphasizes introducing additional intermediate steps to improve final answer accuracy, requiring models to output step-by-step reasoning traces. Research has found that longer CoTs not only provide interpretability but also enhance final answer accuracy, a phenomenon that will be further discussed in Section 4. CoT reasoning fine-tuning uses data in long CoT format (either annotated by human experts or generated synthetically) and applies the same supervised training methodology as instruction tuning to internalize the capability

of producing step-by-step reasoning traces into the model. This approach can also be extended to the multimodal domain. For example, the CoT data in Video-of-Thought [36] divides the process of answering a video QA question into five steps: analyzing the user’s question, constructing a scene graph of the input video, generating detailed video captions, using the acquired information to analyze which option is optimal by comparing against the question and choices, and finally summarizing the entire reasoning process to return the answer.

Video-grounded CoT Fine-tuning. Early CoT reasoning fine-tuning took video and prompts as input and produced pure text as output, which to some extent limited the capabilities of Video-LMMs. Text-only CoTs emphasize logical structure but risk visual hallucination. Therefore, incorporating vision-grounded information into CoTs is beneficial. Video-grounded CoTs reduce hallucination by binding steps to visual evidence via timestamps, shot IDs, or frame indices. VideoEspresso [16] demonstrates that pairing CoT with core frame selection yields fine-grained reasoning supervision while controlling token budgets. ViTCOT [18] advocates video-text interleaving during reasoning, periodically revisiting key frames while thinking, to better align cognition with perception. CoTasks [17] further structures the reasoning interface by injecting entity-level intermediate steps (localization, tracking, relation extraction) as part of the supervision, improving compositional spatiotemporal reasoning.

CoT Fine-tuning for Video RL Cold-Start. Although works such as Video-of-Thought [36] and VideoEspresso [16] have achieved certain success in introducing CoT to video reasoning, the CoT formats used in these datasets are typically fixed, following rigid step sequences. While unified formats facilitate batch generation, they consequently lack flexibility: models cannot explore independently, and predefined paths may not be optimal. Errors generated during fixed-path reasoning cannot be effectively corrected and accumulate continuously. Fundamentally, this represents a static learning paradigm whose effectiveness is highly dependent on the quality and diversity of training data. These models can only imitate the reasoning patterns present in their dataset and struggle to generalize to unseen, more complex scenarios [161]. To overcome this limitation and enable models to learn and align with more abstract and qualitative objectives that are difficult to define precisely in a supervised dataset, many works are increasingly turning to Reinforcement Learning (RL, which will be detailed in), particularly following the emergence of R1-style and GRPO algorithms. Consequently, CoT-SFT has gradually evolved into the cold-start training phase for RL. The cold-start phase is now critical for stabilizing the model before full RL training, preventing instability that can arise from purely RL-driven updates. Cold-start data preparation focuses on capturing human-readable reasoning patterns to prevent instability from purely RL-driven updates. This step generates CoT-style examples with consistent `<think>` and `<answer>` fields, usually involving thousands of carefully curated samples. Structured CoT formats and consistent fields ensure clarity and robustness in the model’s reasoning outputs, reducing errors and improving interpretability [145].

2.3. Data Construction and Representative Resources

Curation Pipelines. Obtaining high-quality video CoT supervision is resource-intensive. A practical approach involves a two-phase curation pipeline: (1) eliciting preliminary CoTs from a reasoning-capable LLM using structured video metadata such as scene descriptions, automatic speech recognition (ASR) transcripts, and shot lists; (2) applying visual consistency refinement through an LMM conditioned on actual video frames to reduce hallucination and align reasoning steps with visual evidence. VideoRFT [60] exemplifies this methodology and provides the VideoRFT-CoT-102K dataset for SFT alongside larger collections designed for RL training.

CoT-SFT Datasets for Video Reasoning. We highlight representative resources used for SFT with CoT format. VideoRFT-CoT-102K supplies large-scale CoT traces tailored for reward-driven fine-tuning and incentivized video reasoning [60]. PixelReasoner-SFT offers pixel/region-grounded, stepwise

supervision that tightly couples perception with structured reasoning. Ego-CoTT-25k targets egocentric and embodied scenarios with chain-of-tool-thought style supervision for ultra-long videos [64]. LongVideo-Reason-CoT [35] extends to multi-event, long-form understanding with narrative-level annotations and supports long-context training pipelines. MTVR-CoT-72k [37], including MTVR-CoT and MTVR-CoT-Tool, contribute multi-task CoT trajectories that bridge video QA and temporal grounding, enabling explicit intermediate reasoning. Beyond the above, fine-grained CoT resources such as VideoEspresso [16], entity-centric CoTasks [17], and interleaved video–text protocols ViT-CoT/ViTIB [18] are widely used as warm-up data, while causal/multi-event understanding can leverage MECD [66]. In addition, Video-of-Thought style collections and their perception-to-cognition protocols provide useful templates for supervising intermediate steps [36]. More resources are summarized in Table 4.

Long-Video Considerations. For long-form video content, SFT typically combines CoT supervision with token-budget control mechanisms, such as shot selection and quota assignment, to maintain computational tractability. These approaches may leverage agentic keyframe selection strategies or frame-aware reasoning signals [34]. When SFT precedes RL training on long videos, as demonstrated in LongVILA-R1 [35], CoT-SFT establishes the format prior that facilitates efficient rollouts and subsequent policy optimization [35].

3. Reinforcement Learning for Video Reasoning

Takeaways

- GRPO has emerged as a popular approach in recent work on video reasoning because it uses verifiable outcomes like answer correctness for optimization, avoiding the need for human preference data.
- A successful system requires co-designing three key elements: advanced policy algorithms, multi-faceted reward functions, and high-quality curated datasets.
- This reinforcement learning approach is highly data-efficient, as a small set of quality data can match or exceed the performance of large-scale supervised tuning.

3.1. Preliminary: From PPO to GRPO

This subsection formalizes three alignment routes that underpin post-training for video reasoning: PPO-based RLHF (with or without AI-generated preferences), Direct Preference Optimization (DPO), and Group Relative Policy Optimization (GRPO). We use x for the multimodal context, y for a response, and τ for a token trajectory.

PPO, RLHF, and RLAIF. RLHF trains a reward model (RM) to score responses and then optimizes the policy with PPO under a KL constraint to a reference model. The RM is commonly trained on preference pairs (x, y^+, y^-) via a Bradley–Terry objective,

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(x, y^+, y^-)} \log \sigma(r_\phi(x, y^+) - r_\phi(x, y^-)),$$

where $r_\phi(x, y)$ is the scalar reward and σ is the logistic function. Given a fixed RM, PPO maximizes a clipped policy-gradient objective with a KL penalty to the reference π_{ref} (e.g., SFT model). Let $r_t(\theta) = \frac{\pi_\theta(y_t | x, y_{\leq t})}{\pi_{\theta_{\text{old}}}(y_t | x, y_{\leq t})}$ and \hat{A}_t be an advantage estimator (often sequence-level reward broadcast to

Table 1 | Summary of large multimodal models for video reasoning (Video-LMMs), including model name, number of parameters, training strategy, test-time scaling, and links.

Model	# Params	# Stages	Training Strategy	TTS	Link
Fact-R1 [87]	~7B	3	SFT + DPO + GRPO	✓	👤 😊
Temporal-RLT [86]	7B	2	SFT + GRPO	✓	👤 😊
VideoChat-R1 [22]	7B	1	Multi-task RFT (GRPO)	✓	👤 😊
Spatial-R1 [84]	7B	1	Task-Specific RFT (GRPO)	✗	👤 😊
LLaVA-NeXT-Video-Thinking [80]	7B - 34B	2	SFT + TDPO (RLHF-style, Segment-Weighted)	✓	👤 😊
video-SALMONN-o1 [79]	7B	2	SFT (LoRA) + pDPO (Process-level)	✓	👤 😊
LongVILA-R1 [35]	7B - 8B	2	CoT-SFT + RL (MR-SP, GRPO)	✗	👤 😊
Video-RTS [23]	7B	1	Pure RL, no SFT (GRPO)	✓	👤 😊
Ego-R1 [64]	~3B	2	SFT (CoTT) + RL (GRPO)	✓	👤 😊
DeepVideo-R1 [81]	2B - 7B	1	Regressive GRPO (Reg-GRPO)	✓	👤 😊
VideoRFT [60]	~7B	2	SFT (CoT) + RL (GRPO)	✓	👤 😊
UniVG-R1 [89]	2B - 7B	2	CoT-SFT + RL (GRPO)	✓	👤 😊
TinyLLaVA-Video-R1 [162]	~3B	2	SFT (Cold-Start) + RL (GRPO)	✓	👤 😊
Video-R1 [21]	7B	2	CoT-SFT + RL (Temporal GRPO)	✓	👤 😊
VAU-R1 [88]	2B - 3B	2	SFT + RFT (GRPO)	✓	👤 😊
ST-R1 [75]	~7B	2	CoT-SFT + RL (GRPO)	✓	👤 😊
TimeZero [96]	~7B	1	Pure RL (GRPO)	✓	👤 😊
VerIPO [25]	7B	3	GRPO-Verifier-DPO loop	✓	👤 😊
VLN-R1 [91]	~7B	2	SFT + RFT (Custom Reward)	✗	👤 😊
TVG-R1 [65]	~7B	2	SFT + RFT	✓	👤 😊
VideoCap-R1 [85]	~7B	2	SFT + RL (GRPO)	✓	-
Vad-R1 [163]	~7B	2	P2C-CoT SFT + AVA-GRPO	✓	👤 😊
R1-SGG [164]	2B - 7B	2	SFT + RL (GRPO)	✗	👤 😊
vsGRPO [83]	2B - 7B	1	R1-Zero-like RL training (GRPO)	✓	👤 😊
BusterX [165]	~7B	2	SFT (Cold-start) + RL (PEFT, DAPO)	✓	👤 😊
ARC-Hunyuan-Video [166]	7B	4	SFT + CoT SFT + RL (GRPO) + SFT	✗	👤 😊
VITAL [37]	7B	7	SFT + Tool-Augmented DGRPO	✓	👤 😊
Video-MTR [78]	~7B	1	RL with Gated Bi-Level Reward (PPO)	✗	- 😊
ReasonAct [74]	3B	3	SFT + V-SFT + Temporal RL (T-GRPO)	✗	-
ReFoCUS [167]	-	2	RL with Reward Model (GRPO)	✗	- 😊
Kwai Keye-VL [168]	8.4B	2	SFT + MPO + Mix-Mode RL (MPO, GRPO)	✓	👤 😊
VRAgent-R1 [90]	-	2	Progressive RL for User Simulation (GRPO)	✓	-
Omni-R1 [169]	7B	2	End-to-End RL (GRPO)	✗	👤 😊
A2Seek-R1 [170]	~3B	2	GoT-SFT + RFT (Aerial GRPO)	✗	👤 -
Pixel Reasoner [115]	7B	2	SFT + Curiosity-Driven RL (Custom)	✓	👤 😊
Tempo-RO [171]	~7B	2	SFT + RFT (PIR-GRPO)	✗	-
VideoSafety-R1 [172]	-	2	AT-SFT + RLHF-style (GRPO)	✓	-
SiLVR [114]	7B - 72B	N/A	Training-Free, Modular	✗	👤 -
CoT-Vid [24]	7B	N/A	Training-Free, Inference-time strategy	✓	-
MR. Video [112]	Modular	N/A	Training-Free, MapReduce Framework	✓	👤 😊
Free-MoRef [113]	7B	N/A	Training-Free, Inference-time MoE	✗	👤 -

tokens):

$$\mathcal{L}_{\text{PPO}}(\theta) = -\mathbb{E} \left[\sum_t \min \left(r_t(\theta) \hat{A}_t, \text{clip} \left(r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right) \right] + \beta \text{KL}(\pi_\theta(\cdot|x) \parallel \pi_{\text{ref}}(\cdot|x)).$$

RLAIF replaces human preferences with AI-generated preferences or rewards; the optimization is unchanged, only the supervision source for \mathcal{L}_{RM} differs. In our curated corpus of video-LLM papers, explicit post-training with PPO/RLHF/RLAIF is uncommon relative to DPO/GRPO.

Direct Preference Optimization (DPO). DPO dispenses with an explicit RM and directly matches the policy to the observed preferences relative to a fixed reference model π_{ref} . With temperature $\beta > 0$, the standard DPO loss over (x, y^+, y^-) is

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E} \log \sigma \left(\beta \left[\log \pi_\theta(y^+|x) - \log \pi_{\text{ref}}(y^+|x) - \log \pi_\theta(y^-|x) + \log \pi_{\text{ref}}(y^-|x) \right] \right).$$

Equivalently, DPO can be viewed as maximizing the log-odds that the policy assigns higher normalized preference to y^+ than to y^- , implicitly inducing a reward proportional to $\log \pi_\theta(\cdot|x) - \log \pi_{\text{ref}}(\cdot|x)$. Recent video-LLMs instantiate this route with process-/task-aware variants, including video-SALMONN-01 (process-DPO) [79], Fact-R1 (preference stage) [87], and LLaVA-NeXTVideo-7B-Thinking (TDPO) [80].

Group Relative Policy Optimization (GRPO). GRPO replaces learned rewards with verifiable outcome rules and optimizes with group-relative advantages. For each prompt x , sample K trajectories $\{\tau^{(k)}\}_{k=1}^K$ from $\pi_{\theta_{\text{old}}}$, compute verifiable scores $r^{(k)} \in [0, 1]$ (e.g., answer correctness, temporal IoU, format checks), and form the group baseline $\bar{r} = \frac{1}{K} \sum_{j=1}^K r^{(j)}$. Define advantages

$$A^{(k)} = r^{(k)} - \text{stopgrad}(\bar{r}), \quad \ell^{(k)}(\theta) = \sum_{t \in \tau^{(k)}} \log \pi_\theta(y_t|x, y_{<t}),$$

and optimize a KL-regularized objective,

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{K} \sum_{k=1}^K A^{(k)} \ell^{(k)}(\theta) + \beta \text{KL}(\pi_\theta(\cdot|x) \parallel \pi_{\text{ref}}(\cdot|x)).$$

In practice, temperature/top- p controls, sequence-length penalties, entropy scheduling, and rejection of malformed traces stabilize on-policy sampling while preserving the verifiable nature of $r^{(k)}$. Recent research have explored GRPO for video-LLMs, including VideoChat-R1 [22], SpaceR [84], Fact-R1 (final RL stage) [87], Reinforcement Learning Tuning for VideoLLMs [86], Scaling RL to Long Videos [35], Video-RTS [23], DeepVideo-R1 [81], Ego-R1 Agent [64], and so on [98, 165, 167, 171, 173–191].

3.2. Video-Specific Policy Optimization

Policy and trajectory formulation. Let $x = (V, q)$ denote the video and query. A trajectory τ interleaves reasoning and decision tokens,

$$\tau = (r_1, \dots, r_{k_1}, d_1, r_{k_1+1}, \dots, d_2, \dots, y),$$

where decisions may propose temporal spans $[t_s, t_e]$, select keyframes \mathcal{F} , emit spatio-temporal regions, and finally produce the answer y . The policy $\pi_\theta(\tau|x)$ factorizes autoregressively. For a group of K rollouts $\{\tau^{(k)}\}$ with verifiable base rewards $r_{\text{base}}^{(k)} \in [0, 1]$, let $\bar{r} = \frac{1}{K} \sum_j r^{(j)}$.

Temporal GRPO (T-GRPO). For each (V, q) , construct two input settings: the ordered frame sequence and a randomly shuffled sequence. Generate two groups of responses and compute the proportions of correct answers p_{ord} and p_{shuf} . Define a temporal coefficient with margin $m \geq 0$:

$$c_{\text{temp}} = \max(0, p_{\text{ord}} - p_{\text{shuf}} - m).$$

For an ordered rollout k , shape the reward

$$r^{(k)} = r_{\text{base}}^{(k)} + \lambda_{\text{temp}} c_{\text{temp}} \mathbb{1}[\text{correct}(\tau^{(k)})],$$

and set the group-relative advantage $A^{(k)} = r^{(k)} - \bar{r}$. The GRPO update maximizes $\sum_k A^{(k)} \sum_{t \in \tau^{(k)}} \log \pi_{\theta}(y_t | x, y_{<t})$ under a KL anchor to π_{ref} , which explicitly rewards accuracy that depends on temporal order rather than single-frame shortcuts [21].

Regressive GRPO (Reg-GRPO). Reg-GRPO [81] reformulates GRPO as regression on group-normalized advantages, removing min/clipping safeguards. Let the normalized target be

$$\tilde{A}^{(k)} = \frac{r_{\text{base}}^{(k)} - \mu_r}{\sigma_r}, \quad \mu_r = \frac{1}{K} \sum_j r_{\text{base}}^{(j)}, \quad \sigma_r = \sqrt{\frac{1}{K} \sum_j (r_{\text{base}}^{(j)} - \mu_r)^2}.$$

Define a sequence score $s_{\theta}(\tau^{(k)}, x) = \sum_{t \in \tau^{(k)}} \log \pi_{\theta}(y_t | x, y_{<t})$. The loss is

$$\mathcal{L}_{\text{Reg-GRPO}}(\theta) = \frac{1}{K} \sum_{k=1}^K \left(s_{\theta}(\tau^{(k)}, x) - \tilde{A}^{(k)} \right)^2 + \beta \text{KL}(\pi_{\theta}(\cdot | x) \| \pi_{\text{ref}}(\cdot | x)).$$

To mitigate vanishing advantages on very easy/hard samples, DeepVideo-R1 adds difficulty-aware augmentation and/or per-sample weights $w(d(x))$:

$$\mathcal{L}_{\text{Reg-GRPO}}^{\text{DA}}(\theta) = \frac{1}{K} \sum_k w(d(x)) \left(s_{\theta}(\tau^{(k)}, x) - \tilde{A}^{(k)} \right)^2 + \beta \text{KL}(\cdot).$$

Token-weighted advantages (TW-GRPO). To improve credit assignment along long chains of thought, TW-GRPO introduces token importance w_t estimated from intra-group informativeness (e.g., entropy across the K rollouts). Replace the unweighted score with

$$s_{\theta}^{\text{TW}}(\tau^{(k)}, x) = \sum_{t \in \tau^{(k)}} w_t \log \pi_{\theta}(y_t | x, y_{<t}),$$

and compute advantages from a soft multi-bin reward $r_{\text{soft}}^{(k)} = \sum_b \gamma_b \mathbb{1}[y^{(k)} \in \mathcal{Y}_b]$ (exact, near-miss, wrong). The resulting GRPO/Reg-GRPO objective uses s_{θ}^{TW} in place of s_{θ} , yielding denser, lower-variance updates [82].

Difficulty-aware GRPO (DGRPO). To address the difficulty imbalance across tasks or prompts, DGRPO reweights the group-relative advantages by adaptive difficulty signals. Let d_{task} be a moving hardness estimate at the task level and d_{sample} a per-prompt score (e.g., running success rate or verifier score dispersion). With a monotone weight $g(\cdot, \cdot)$,

$$\tilde{A}_{\text{DA}}^{(k)} = g(d_{\text{task}}, d_{\text{sample}}) (r^{(k)} - \bar{r}),$$

and the update maximizes $\sum_k \tilde{A}_{\text{DA}}^{(k)} \sum_{t \in \tau^{(k)}} \log \pi_{\theta}(y_t | x, y_{<t})$ under the same KL anchor. In “Thinking With Videos,” this scheme is used together with curated multi-task RL data (MTVR-RL-110k) to emphasize informative failures and prevent easy examples from dominating [37].

Table 2 | Policy optimization methods for Video-LMM post-training. GRPO-family, preference-based alignment, verifier-guided pipelines, and long-video variants.

Method	Objective	Symbols
Vanilla GRPO [22, 86]	$\max_{\theta} \frac{1}{G} \sum_{i=1}^G \min\left(\frac{\pi_{\theta}(y_i)}{\pi_{\text{old}}(y_i)} A_i, \text{clip}\left(\frac{\pi_{\theta}}{\pi_{\text{old}}}, 1-\epsilon, 1+\epsilon\right) A_i\right) - \beta \text{KL}[\pi_{\theta} \parallel \pi_{\text{ref}}]$	G : group size; $A_i = \frac{r_i - \mu_r}{\sigma_r}$; r_i : verifiable reward; π_{ref} : reference policy; $\text{clip}(\cdot)$: PPO-style clipping
T-GRPO [21]	$\max_{\theta} \mathcal{L}_{\text{GRPO}}(\theta) + \lambda_t \alpha \mathbf{1}[p_{\text{ord}} > p_{\text{shuf}}]$	$p_{\text{ord}}, p_{\text{shuf}}$: success on ordered vs. shuffled frames; λ_t, α : weights
TW-GRPO [82]	$\max_{\theta} \frac{1}{G} \sum_i \min\left(\frac{\pi_{\theta}}{\pi_{\text{old}}} A'_i, \text{clip}(\cdot) A'_i\right) - \beta \text{KL}, \quad A'_i = \sum_t w_t a_{it}, \quad r = \sum_k \gamma_k \mathbf{1}[y \in \mathcal{Y}_k]$	A'_i : token-weighted advantage; w_t : token importance; a_{it} : token-level advantage; \mathcal{Y}_k : partial-credit bins; γ_k : bin weights
Reg-GRPO [81]	$\min_{\theta} \frac{1}{G} \sum_{i=1}^G \left(\Delta \log \pi_{\theta}(y_i) - \eta A_i \right)^2 + \beta \text{KL}$	$\Delta \log \pi_{\theta}(y_i) = \log \frac{\pi_{\theta}(y_i)}{\pi_{\text{old}}(y_i)}$; η : regression scale
DGRPO [87]	$\max_{\theta} \frac{1}{G} \sum_i \min\left(\frac{\pi_{\theta}}{\pi_{\text{old}}} A''_i, \text{clip}(\cdot) A''_i\right) - \beta \text{KL}, \quad A''_i = w(d(x)) \cdot A_i$	$d(x)$: difficulty score; $w(\cdot)$: difficulty weight; A''_i : difficulty-weighted advantage
Multi-task GRPO [22]	$\max_{\theta} \frac{1}{G} \sum_i \min\left(\frac{\pi_{\theta}}{\pi_{\text{old}}} \sum_m \lambda_m A_i^{(m)}, \text{clip}(\cdot) \sum_m \lambda_m A_i^{(m)}\right) - \beta \text{KL}$	$A_i^{(m)}$: standardized advantage on task m ; λ_m : task weights
Verifier-DPO [25]	$\min_{\theta} \mathcal{L}_{\text{DPO}} = -\log \frac{\exp(\beta s^+)}{\exp(\beta s^+) + \exp(\beta s^-)}$	s^+, s^- : scores of preferred/rejected outputs; β : DPO temperature
Long-video-RL [35]	$\max_{\theta} \sum_{s=1}^S \mathcal{L}_{\text{GRPO}}^{(s)}(\theta) - \gamma \Omega(\text{memory/retrieval})$	S : #segments; $\mathcal{L}_{\text{GRPO}}^{(s)}$: per-segment objective; $\Omega(\cdot)$: memory/retrieval regularizer; γ : weight
Temporal-only grounding RL [86]	$\max_{\theta} \mathcal{L}_{\text{GRPO}}(\theta) \text{ s.t. } r = \text{IoU}([t_s, t_e], [\hat{t}_s, \hat{t}_e])$	$[t_s, t_e]$: predicted span; $[\hat{t}_s, \hat{t}_e]$: ground-truth span
Spatio-temporal GRPO [22]	$\max_{\theta} \mathcal{L}_{\text{GRPO}}(\theta) \text{ with } r = \lambda_f R_{\text{format}} + \lambda_{\text{IoU}} R_{\text{IoU}} + \lambda_a R_{\text{acc}} + \lambda_r R_{\text{recall}}$	R_{format} : structured output; R_{IoU} : temporal IoU; R_{acc} : MC/classification accuracy; R_{recall} : event recall
Caption w/ dual verifiable rewards [85]	$\max_{\theta} \mathcal{L}_{\text{GRPO}}(\theta) \text{ with } r = \lambda_f R_{\text{format}} + \lambda_c R_{\text{content}}$	R_{format} : template/structure score; R_{content} : content fidelity; λ_f, λ_c : weights
RLxRTS [23]	$\max_{\theta} \mathcal{L}_{\text{GRPO}}(\theta) - \lambda_s \Phi(\text{CoT steps})$	λ_s : coupling weight; $\Phi(\cdot)$: penalty/constraint on CoT step count

Note: DPO can be viewed as an offline preference-based alignment method related to RL.

3.3. Reward Design for Video Reasoning

We decompose the outcome reward into verifiable components and aggregate them with task weights:

$$R(x, \tau) = \sum_m \lambda_m R_m(x, \tau), \quad \lambda_m \geq 0, \quad \sum_m \lambda_m = 1,$$

Table 3 | Reward design taxonomy for Video-LMM post-training.

Aspect	Typical formulation	Examples
Temporal localization	Span IoU/mIoU; event order consistency; count/duration constraints	[21, 22, 86]
Spatial grounding	Box/mask/track IoU; trajectory overlap; relation/pose consistency	[22, 84]
Content correctness	MC accuracy; open-ended semantic match; partial-credit bins	[82, 87]
Format/structure	Enforce <code><think>/<answer></code> template; reasoning-step completeness	[86, 87]
Hallucination mitigation	Entity/evidence grounding checks; cross-modal consistency penalty	[80, 128]
Difficulty-aware weighting	$w(d(x))$ on advantages; curriculum by hardness bins	[37]
Tool-augmented signals	Reward for informative frame retrieval; toolbox success/failure	[37, 64]
Memory/retrieval regularization	Penalty $\Omega(\cdot)$ on memory calls; segment-wise consistency	[35]
Audio-aware consistency	Optional ASR/AV alignment scores when audio is used	[79]

which distributes incentives and mitigates reward hacking by avoiding reliance on any single objective.

Format and faithfulness. Outputs are parsed with lightweight rules (e.g., required `<think>/<answer>` tags, unit normalization, timestamp presence, citation syntax). Violations incur graded penalties; contradictions with visual or subtitle evidence trigger additional deductions [25].

Answer correctness. For multiple-choice, we use exact match. For open-ended responses, we compute normalized string scores (e.g., edit distance, token-F1) with minor lexical normalization and, when necessary, a calibrated evaluator to assign partial credit rather than binary pass/fail [60].

Temporal localization. Given a predicted interval $P = [t_s, t_e]$ and ground truth G , we combine smooth temporal IoU and threshold bonuses while discouraging overlong spans:

$$R_{\text{temp}} = \alpha \text{tIoU}(P, G) + \sum_k b_k \mathbb{1}[\text{tIoU}(P, G) \geq \tau_k] - \gamma \frac{|P|}{|V|}.$$

Missed critical events (false negatives) receive additional penalties to avoid degenerate short spans [27, 28].

Spatio-temporal grounding. For regions or tracks $\{B_t\}$, we combine region-IoU/track-IoU with center-distance shaping and enforce text-region referential consistency across frames to prevent hallucinated references [22].

Budget awareness. Let B be the frame/token budget. We reward accurate solutions that respect B and penalize redundant re-observations; staged viewing (coarse-to-fine frame selection) receives a small bonus:

$$R_{\text{budget}} = \eta_1 \mathbb{1}[\text{correct}] \cdot \left(1 - \frac{\text{used}}{B}\right) - \eta_2 \frac{\text{repeats}}{\text{used}}.$$

This keeps the policy sample-efficient during long-video rollouts [30, 31].

Verifier and critic signals. External verifiers check timestamp/region claims and entity references; multi-path self-consistency (e.g., majority vote or agreement rate across K sampled traces) yields pass/fail or graded signals that fold into R and help stabilize exploration [25, 37].

Aggregation and normalization. Task weights $\{\lambda_m\}$ are tuned to equalize gradient magnitudes across objectives. We normalize each R_m to $[0, 1]$ on a per-batch basis and apply temperature scaling when mixing discrete pass/fail terms with continuous IoU-style signals. This keeps the GRPO advantages well-conditioned and reduces variance during on-policy sampling.

3.4. RL Datasets for Video Reasoning

Reinforcement learning for video reasoning draws on three complementary data sources. First, supervised chain-of-thought corpora warm up the policy to produce structured traces that can be scored online by verifiers. Second, RL rollout corpora provide prompts with verifiable targets, e.g., answer strings, timestamps, or regions, so that outcome rewards can be computed without human preferences. Third, curated hard negatives and near-duplicate distractors sharpen temporal and spatial discrimination under limited budgets.

Representative scales and staging. Across recent Video-LMMs the RL data footprint ranges from a few thousand to hundreds of thousands of examples, often after a smaller SFT warmup. Video-RTS demonstrates a single-stage GRPO pipeline trained on roughly 6K video-QA triples, yet matches systems that rely on ~ 165 K SFT pairs, highlighting data efficiency under verifiable rewards [23]. LongVILA adopts a two-phase schedule: long-video CoT-SFT on about 36K samples, followed by GRPO with ~ 68 K filtered prompts plus ~ 102 K external additions to stabilize exploration at length [35]. Fact-R1 explicitly separates stages, ~ 85 K long-form CoT-SFT, then ~ 5 K preference pairs for DPO alignment, and finally GRPO with verifier-backed outcome rewards while jointly training auxiliary caption/OCR heads [87]. Multi-task GRPO in VideoChat-R1 operates over a mixed training set totaling approximately 18,031 samples spanning QA, grounding, tracking, and captioning, showing that a moderate-scale, heterogeneous pool suffices when rewards are verifiable [22]. Larger pipelines exist as well: ARC-Hunyuan-Video-7B [166] reports instruction-tuning corpora on the order of 4.6×10^5 pairs and tens of thousands of GRPO rollouts distributed across tasks, interleaved with cold-start and polish stages to control drift.

Temporal and spatial supervision. Effective RL corpora emphasize prompts with temporal anchors and spatial references so that rewards can combine correctness with localization. Typical sources include timestamped QA, dense event or action segments, and region-grounded queries. For long-form content, authors construct silver labels with shot detection and ASR alignment to produce answerable windows and span-level targets, which enable smooth tIoU shaping during GRPO [23, 35].

Curation and filtering. To control reward hacking and variance, recent works filter prompts for unambiguous answers, enforce strict formatting constraints, and mine hard negatives from near-duplicate shots or distractor spans before rollout. In practice this yields a compact but high-yield RL pool (e.g., the ~ 68 K filtered set in LongVILA) that keeps the verifier precise and the advantages well-conditioned [35].

Domain breadth and streaming settings. Beyond general video QA, RL datasets extend to navigation, egocentric, and streaming regimes where budgets and latency matter. For example, StreamVLN trains over hundreds of thousands of trajectories and on the order of 6×10^7 frames with a GRPO-style objective adapted to streaming perception and action, illustrating how outcome rewards transfer to embodied video tasks [192].

4. Test-Time Scaling for Video Reasoning

Takeaways

Test-time scaling improves reliability by allocating inference compute to evidence selection, reasoning depth, and path diversity. Recent work has explored various TTS strategies, including Video-CoT prompting, self-consistency with verifier gating, confidence-guided iteration with refine-on-fail, and tool-augmented chains for long or streaming videos.

4.1. Beam Search for Video Outputs

Beam search is a standard decoding strategy adopted by many video captioning and video-QA models to improve the fluency and relevance of generated text. In video captioning tasks, for example, models often generate descriptions using beam search (e.g., beam width 5) to explore multiple candidate sentences and pick the best one. This approach has been used to produce higher-quality captions by balancing completeness and coherence as compared to greedy decoding. Overall, beam search serves as a test-time decoding boost for Video-LMMs by considering alternative word sequences and selecting the highest-probability caption.

4.2. Video Chain-of-Thought Prompting

CoT prompting, getting the model to generate intermediate reasoning steps before the final answer, has been successfully extended to video understanding. Video-of-Thought (VoT) [36] was one of the first frameworks to implement CoT for video reasoning. VoT [36] breaks a complex video question into simpler sub-problems and addresses them step by step, from low-level perceptual cues to high-level conclusions. This explicit reasoning significantly improved performance on challenging video QA benchmarks, demonstrating the benefit of prompted reasoning traces in video tasks. More recently, CoT-Vid [24] introduced a training-free multi-stage CoT pipeline for video QA. CoT-Vid [24] dynamically decides whether a question needs reasoning, then decomposes it and iteratively reasons step by step before producing the answer, yielding notable accuracy gains without any model fine-tuning.

4.3. Self-Consistency Decoding in Video Reasoning

Video-LMMs have also begun to employ self-consistency decoding, where multiple reasoning paths are sampled and then aggregated to improve answer reliability. A clear example is the video self-consistency verification stage in CoT-Vid [24]. During inference, CoT-Vid [24] generates multiple reasoning chains for the same question and uses a similarity-based voting mechanism to merge them into a final answer. This ensures that the chosen answer is consistent with the majority of reasoning paths and with the video content, reducing random errors or hallucinations. Empirically, video self-consistency yields better accuracy as more answer samples are considered, CoT-Vid's performance improved steadily up to about five reasoning samples before saturating, stabilizing outputs by leveraging ensemble reasoning.

4.4. Confidence-Based Iterative Reasoning

Recent Video-LMM agents use confidence measures to guide and terminate multi-step inference. CyberV [108] treats reasoning as a closed-loop process: a controller monitors uncertainty and instructs the model to think deeper or request denser visual evidence until a stopping criterion is met. Video-

ICL [109] similarly allocates more computation to uncertain queries and stops early on confident ones. This confidence-driven iteration allows Video-LMMs to balance thoroughness and efficiency by refining their understanding progressively and stopping only when the answer is likely correct.

4.5. Self-Improvement via Refinement Loops

Several video reasoning frameworks implement iterative self-refinement loops at test time, enabling the model to improve answers over multiple rounds. DIVE (Deep-search Iterative Video Exploration) [110] breaks down each question into sub-questions and tackles them in a multi-step loop, refining the queries and answers at each pass. If an intermediate answer is incomplete or a sub-question remains, DIVE [110] re-evaluates and refines that part in the next iteration. This refine-on-fail strategy yields highly accurate and contextually appropriate answers even for complex queries. Similarly, Video-MTR [78] performs multi-turn reasoning on long videos, progressively selecting relevant segments and updating the answer until convergence.

4.6. Monte Carlo Tree Search (MCTS) for Video-LMMs

Monte Carlo Tree Search has been applied to expand and diversify generation at inference. Auto-Caption [111] uses MCTS to iteratively construct diverse video descriptions by exploring a tree of possible continuations and selecting branches that yield informative sentences. This produces rich sets of key-point captions that go beyond fixed-beam decoding, and enables the MCTS-VCB benchmark where MLLMs fine-tuned on AutoCaption outputs show large gains.

4.7. Chain-of-Action and Tool-Augmented Reasoning

Video-LMMs are increasingly embracing tool use and multi-step action chains to handle complex video understanding. VITAL [37] equips a video-language model with a visual toolbox that the model can call during reasoning. At inference time, VITAL [37] decides when to invoke tools (for example, to fetch a particular video clip segment or detect an object) and incorporates the results into a multimodal chain of thought, greatly reducing hallucinations by grounding intermediate claims in returned evidence. Ego-R1 [64] introduces a Chain-of-Tool-Thought paradigm for ultra-long egocentric videos: an RL-trained agent orchestrates specialized tools in sequence, first calling a temporal retrieval tool to find a relevant moment, then an object recognizer, and so on, each tool tackling a sub-task of the query, enabling answers about weeks-long recordings beyond raw context limits. ReAgent-V [116] coordinates multiple specialized agents and tools so that perception and reasoning are scheduled and verified under long or streaming inputs. Complementary agentic strategies include VideoDeepResearch [107], which performs tool-augmented search over long videos at inference time, and Agentic Keyframe Search [103], which plans which frames to inspect and couples planner-executor loops with verification before answer commitment.

5. Benchmarks for Video-LMM Post-training Evaluation

Evaluating post-training requires benchmarks aligned with optimization objectives: verifiable supervision for RL, realistic compute budgets for TTS, and protocols that expose genuine reasoning rather than shortcut exploitation. We organize resources into general QA, video reasoning, and grounding-centric benchmarks, emphasizing settings that enable verifier-ready rewards and standardized comparisons. Table 4 summarizes commonly used datasets in recent post-training work.

Table 4 | Datasets used in Video-LLM post-training (training & evaluation). Row color indicates primary usage scenario, and datasets may be used across multiple stages: **SFT**, **RL**, **Bench**.

Name (with source)	Size	Tasks	Link
Temporal-RLT-Full-490k [86]	490,000	VideoQA, temporal grounding, grounded VideoQA; diversified difficulty; used before RL.	🤗
Temporal-RLT-32k [86]	32,000	Curated subset for GRPO-style RLT; temporal signals emphasized.	🤗
VideoChat-R1 training set [22]	18,031	Multi-task SFT covering grounding, tracking, grounded QA.	—
MTVR-CoT-72k [37]	72,000	Long CoT reasoning; temporal grounding; tool-augmented SFT variants included.	🤗
MTVR-RL-110k [37]	110,000	Multi-task video reasoning; difficulty-aware scheduling.	🤗
Video-R1-COT-165k [21]	165,000	Chain-of-thought supervision for time-aware reasoning (ordered vs. shuffled frames).	🤗
Video-R1-260k [21]	260,000	RL pool for T-GRPO reinforcement; mixed video/image subsets.	🤗
video-SALMONN-o1 (QA pairs) [79]	~180,000 QA (from ~13k videos)	Audio+video reasoning; curated QA pairs for instruction/SFT.	—
video-SALMONN-o1 (preferences) [79]	~200,000 pairs	Pairwise preference data for DPO/RFT-like objectives; strengthens chain-of-thought quality.	—
LongVILA CoT-SFT [35]	36,000	Long-video chain-of-thought supervision; length-aware prompts.	🤗
LongVILA RL pool [35]	68,000 + 102,000 (open)	Two-part RL data (in-house + open-source) targeting long temporal reasoning.	🤗
FakeVV (news-domain) [87]	197,600	Video misinformation detection/explanation; reasoning traces.	👤
FakeTT (short-video, EN) [87]	—	Short-video misinformation (English); used for SFT and analysis.	👤
FakeSV (short-video, ZH) [87]	18,859	Short-video misinformation (Chinese); reasoning.	👤
TVG-Coldstart-13K [65]	~13k	SFT cold-start for temporal grounding	🤗
TVG-RL-18K [65]	~18k	RL data for temporal grounding	🤗
Charades-STA [22]	—	Temporal grounding benchmark.	🤗
ActivityNet-Grounding [22]	—	Temporal grounding benchmark.	👤
ActivityNet-RTL [22, 86]	—	Reasoning-intensive temporal grounding benchmark.	🤗
AVE-2 [193]	570,138	Audio-visual alignment evaluation reasoning.	🤗
GoT-10k [22]	—	Object tracking benchmark.	🤗
NExT-GQA [22]	—	Video QA / grounded QA benchmark.	🤗
Dream-1k [22]	—	Captioning benchmark (dense descriptions).	🤗
VidTAB-QA [22]	—	Video QA quality assessment benchmark.	🤗
VSI-Bench [84]	—	Spatial reasoning (relations, order, counting).	🤗
VideoMME [22]	2,700 QA	General video understanding benchmark.	🤗
MVBench [22]	—	General video understanding benchmark.	🤗
Video-Holmes [194]	—	Video reasoning benchmark.	🤗
MMVU [118]	3,000 items	Expert-level multidisciplinary video.	🤗
Video-MMMU [195]	900 QA pairs	Multi-discipline professional videos.	🤗
VideoHallucer / HAVEN [80]	6,497 QA (HAVEN)	Hallucination evaluation (object/temporal consistency).	🤗

Takeaways

Alignment between evaluation metrics and training objectives enables more interpretable optimization: answer faithfulness, temporal correctness, and spatial-temporal grounding under realistic budgets with verifier-ready annotations. The field has moved beyond monolithic QA suites toward targeted evaluations, including multi-event reasoning, long-video and streaming, and precise grounding, that better diagnose where post-training gains come from.

5.1. General Video QA Benchmarks

Comprehensive QA suites probe recognition, reasoning, and instruction following across diverse lengths and domains [121, 195, 196]. MMVU [118] targets expert-level, multi-discipline understanding and provides dual reporting protocols (with and without subtitles) to expose text-based shortcuts. VCR-Bench [125] focuses on compositional, causal, and multi-step reasoning with fine-grained categories for capability analysis. VideoReasonBench [126] emphasizes vision-centric reasoning beyond frame-level recognition, stressing cross-event inference and temporal dependencies. MINERVA [127] stresses complex multi-step reasoning over long videos, assessing sustained attention and multi-hop inference. Standard metrics include accuracy for multiple-choice and exact match or F1 for free-form answers, with recommended dual reporting with and without subtitles to reveal linguistic shortcut exploitation [118, 125].

5.2. Video Reasoning Benchmarks

Reasoning-centric evaluations isolate capabilities that post-training often targets. MECD [66] measures multi-event causal dependencies, enabling analysis of causal chains across shots. VidHalluc [128] and HAVEN [80] probe hallucination robustness, including temporal hallucination and object consistency, testing whether models fabricate non-existent entities or events. Long-video and streaming settings such as LongVideo-Reason-eval [35] and streaming/multi-round evaluations (e.g., StreamBench [32], SVBench [133], OmniMMI [135]) stress memory management, budgeted viewing, and stability under temporal resampling. For these protocols, budget- and latency-aware reporting is essential: disclose viewing budget (frames or tokens), reasoning length, path count, and latency/throughput alongside accuracy to reveal cost-performance trade-offs critical for deployment [35].

5.3. Grounding Reasoning Benchmarks for Video-LMMs

Grounding-centric benchmarks align tightly with verifiable rewards used in RL and with inference-time verification. Temporal localization datasets such as Charades-STA and ActivityNet Grounding [22] evaluate precise moment retrieval from language, while ActivityNet-RTL [22, 86] requires multi-step reasoning before localization. The fine-grained 0–10 scores make it a verifier-ready resource for RL-based post-training and a bridge between moment-localization benchmarks and multimodal reasoning suites. Spatial-temporal grounding benchmarks broaden the target to regions and tracks: V-STAR [29] provides entity/action grounding with trajectory annotations; VSI-Bench [84] probes spatial relations, ordering, and counting; GoT-10k [130] stresses long-term identity maintenance via object tracking. Evaluation commonly reports temporal IoU (tIoU), Recall@K at multiple tIoU thresholds (e.g., 0.3/0.5/0.7), region/trajectory IoU, and center-distance errors, with locate-then-answer protocols that require models to commit to evidence before producing answers [27? , 28].

5.4. Long and Streaming Video Evaluation

Long/streaming evaluations target long-horizon reasoning, dialogue coherence, and timestamp sensitivity under online constraints. SVBench [133] uses temporally linked multi-turn QA chains to probe streaming understanding; StreamBench [32] evaluates real-time, interactive scenarios. OVO-Bench [197] stresses timestamp-aware online reasoning with three settings, backward tracing, real-time comprehension, and forward (delayed) answering, paired with fine-grained temporal annotations. For long-form video, LongViTU [198] supplies large-scale long-video QA with explicit timestamps, and HLV-1K [131] focuses on hour-long videos. For captioning, AuroraCap [199] introduces VDC (a detailed video captioning benchmark) and VDCscore, an LLM-assisted metric that decomposes long captions into QA-style checks.

6. Challenges and Future Directions

We highlight challenges and promising forward paths that connect SFT, RL, and TTS for video LMMs with a focus on verifiability, efficiency, and robustness. Rather than treating these paradigms as isolated techniques, the field is moving toward deep integration that converts training-time investment into dependable test-time accuracy while addressing concrete limitations reported across recent studies.

Takeaways

- Ground supervision and evaluation in structured, evidence-linked reasoning and explicit verifier signals; actively diagnose and mitigate sycophancy, judge and length biases, and subtitle leakage.
- Scale RL on long videos with verifiable, compositional rewards, efficient frame selection and caching, and exploration objectives that go beyond distilled teachers.
- Build budget-aware anytime agents that couple confidence estimates with verifier checks and tool use; standardize reporting (viewing budget, reasoning length, paths, latency/throughput, subtitle usage) to ensure fair comparison and avoid leakage.

6.1. Future Directions for Video-LMM Supervised Fine-Tuning

Structured interfaces and grounded CoT. Codifying reasoning formats that bind steps to evidence (timestamps, frame IDs, regions) can improve faithfulness and simplify verifier design, building on multimodal CoT resources [16–18]. Normalizing tags, citations, and unit conventions enables plug-and-play checks later used in RL and TTS.

Verifier-in-the-loop CoT synthesis at scale. Automate draft–refine–audit pipelines that start from ASR/OCR/shot metadata, refine on frames, and filter with lightweight checkers to reduce hallucinations. Reduce template and single-model biases by mixing trace generators and including self-correction exemplars; couple instruction tuning to task metrics rather than style alone [26, 106, 172].

Trimodal supervision and subtitle controls. Many queries hinge on audio cues and speaker turns. Extend SFT to align speech, events, and visual evidence and always report with and without transcripts to avoid shortcircuiting via ASR. Current works highlight limited audio coverage and the need for streaming-aware alignment [35, 79, 133].

Hallucination-aware instruction tuning. Incorporating counterfactual and absence cases from robustness resources [80, 128] teaches calibrated abstention and verification-seeking behavior, reducing over-affirmation as chains lengthen.

Multilingual, OCR, and narrative structure. Data remains imbalanced across languages and misses hard OCR and narrative dependencies. Future SFT should target multilingual breadth, degraded text, and long-span story reasoning so improvements transfer beyond narrow scenarios [200, 201].

6.2. Challenges and Future Directions of RL for Video Reasoning

Compositional, verifiable rewards. Beyond tIoU/IoU, many tasks require joint time–space–semantics checks (entity linking, ordering, object–action binding) [22, 27, 28]. Process Reward Models (PRMs) can provide dense credit along chains but need cost-effective construction and bias control [35]. Lightweight rule systems like VeriPO complement PRMs and transfer to TTS verification [25].

Sample efficiency and long-video cost. Caching visual features and decoupled encoders help [35], yet scaling RL still strains budgets. Off-policy and model-based variants, world models, and micro-rollouts (optimize locate-first, then answer) are promising for exploration efficiency [202]. Architectural context-scaling offers another path. For instance, VideoNSA [203] applies learnable, hardware-aware native sparse attention [204], reliably scaling to 128K tokens and improving temporal reasoning over compression-based baselines; MovieChat+ [205] uses question-aware sparse memory to support long-video reasoning without external temporal modules while cutting cost.

Exploration beyond teachers. Curriculum and teacher distillation mitigate cold starts [35], but discovering strategies surpassing teachers requires diversity-driven objectives and self-play. Difficulty-aware and group-relative schemes from recent RL for video provide practical starting points [37].

Evaluation bias and fair comparison. Judge bias and length bias can distort progress when using LLMs as evaluators. Report matched budgets, control for reasoning length, and include human or verifier-based audits to ensure reliability [118, 206].

Scaling beyond preference data. Automated pipelines [60] and self-alignment [207] reduce annotation dependence but must broaden coverage for causal and counterfactual reasoning and diverse domains [20, 86].

6.3. Video-LMM Test-Time Scaling Future Directions

Confidence-aware, verifier-guided TTS. Stopping rules tied to uncertainty, coupled with verifier checks, can deliver anytime accuracy: deepen reasoning or densify viewing only when needed, echoing closed-loop designs and sparse-to-dense schedules [23, 108].

Tool-augmented inference and distillation. Reasoning that interleaves tool calls (retrieval, tracking, ASR alignment) improves faithfulness at test time [37]; post-hoc distillation can transfer these benefits into base models to cut inference cost, using verifier-anchored traces as supervision [25].

Streaming agents with memory. Agentic planners that decide what to watch next and when to stop, while maintaining task-aware working memory, are essential for long or streaming video [32, 33, 208]. Budget-aware rewards can train these behaviors for robust anytime performance.

Standardized reporting and leakage control. Report viewing budgets, reasoning lengths, path counts, latency/throughput, and subtitle usage. Include sycophancy and judge-bias diagnostics so gains are attributable and not artifacts of prompt length or transcript leakage [26, 118].

Compute–accuracy trade-offs under constrained viewing. Co-tune frame selection and compression with reasoning quality so systems remain strong when only a small fraction of frames are processed. Frame-optimization and compression frameworks still incur notable cost; future work should make these components data- and compute-efficient [167, 209].

7. Conclusion

This survey has systematically analyzed the critical role of post-training in advancing video reasoning, tracing the evolution from foundational Supervised Fine-tuning with Chain-of-Thought to more powerful and autonomous paradigms. Reinforcement learning, primarily through online frameworks like GRPO, has become a core engine for optimization, while emerging agentic frameworks and test-time scaling strategies offer new frontiers in reasoning capability and efficiency. Despite these significant advances, the path to robust, general-purpose video intelligence is still marked by key challenges. The future research agenda will be defined by overcoming data scarcity for complex reasoning, developing more sample-efficient and stable RL algorithms, strengthening multimodal grounding to prevent hallucinations, and creating integrated frameworks that synergize training-time alignment with inference-time computation. Addressing these interconnected challenges is crucial to advancing the boundaries of video understanding systems.

References

- [1] Gary Snyder. *The practice of the wild: Essays*. Catapult, 2020.
- [2] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- [3] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [5] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [7] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey, 2025.
- [8] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2025.
- [9] Heqing Zou, Tianze Luo, Guiyang Xie, Victor, Zhang, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, and Huaijian Zhang. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding, 2024.

- [10] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.
- [11] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.
- [12] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021.
- [13] Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Soo Ye Kim, Zhifei Zhang, Yilin Wang, Jianming Zhang, Zhe Lin, and Jiebo Luo. Finecaption: Compositional image captioning focusing on wherever you want at any granularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24763–24773, 2025.
- [14] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025.
- [15] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces, 2025.
- [16] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection, 2024.
- [17] Yanan Wang, Julio Vizcarra, Zhi Li, Hao Niu, and Mori Kurokawa. Cotasks: Chain-of-thought based video instruction tuning tasks, 2025.
- [18] Yongheng Zhang, Xu Liu, Ruihan Tao, Qiguang Chen, Hao Fei, Wanxiang Che, and Libo Qin. Vtcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models, 2025.
- [19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [20] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models, 2025.
- [21] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025.

- [22] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning, 2025.
- [23] Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning, 2025.
- [24] Hongbo Jin, Ruyang Liu, Wenhao Zhang, Guibo Luo, and Ge Li. Cot-vid: Dynamic chain-of-thought routing with self verification for training-free video reasoning, 2025.
- [25] Yunxin Li, Xinyu Chen, Zitao Li, Zhenyu Liu, Longyue Wang, Wenhan Luo, Baotian Hu, and Min Zhang. Veripo: Cultivating long reasoning in video-lmms via verifier-guided iterative policy optimization, 2025.
- [26] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.
- [27] Ziqiang Xu, Qi Dai, Tian Xie, Yifan Yang, Kai Qiu, DongDong Chen, Zuxuan Wu, and Chong Luo. Viarl: Adaptive temporal grounding via visual iterated amplification reinforcement learning, 2025.
- [28] Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding, 2025.
- [29] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-lmms on video spatio-temporal reasoning, 2025.
- [30] Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, Sungjin Ahn, Jan Kautz, Hongxu Yin, Yao Lu, Song Han, and Wonmin Byeon. Storm: Token-efficient long video understanding for multimodal lmms, 2025.
- [31] Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-lm based video frame selection for efficient video understanding, 2025.
- [32] Haomiao Xiong, Zongxin Yang, Jiazu Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge, 2025.
- [33] Zhuo Zhi, Qiangqiang Wu, Minghe Shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. Videoagent2: Enhancing the lm-based agent system for long-form video understanding by uncertainty-aware cot, 2025.
- [34] Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal lmms via frame-aware reasoning, 2025.

- [35] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos, 2025.
- [36] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition, 2024.
- [37] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning, 2025.
- [38] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding, 2025.
- [39] Yogesh Kumar. Videollm benchmarks and evaluation: A survey, 2025.
- [40] Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2025.
- [41] Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.
- [42] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- [43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [44] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.
- [45] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [46] Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. Pave: Patching and adapting video large language models, 2025.
- [47] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024.
- [48] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos, 2024.
- [49] Siddhant Bansal, Michael Wray, and Dima Damen. Hoi-ref: Hand-object interaction referral in egocentric vision. *arXiv preprint arXiv:2404.09933*, 2024.
- [50] Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 7293–7301, 2025.

- [51] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding, 2024.
- [52] Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *European Conference on Computer Vision*, pages 166–185. Springer, 2024.
- [53] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.
- [54] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.
- [55] Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 3599–3607, 2025.
- [56] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.
- [57] Tao Chen, Enwei Zhang, Yuting Gao, Ke Li, Xing Sun, Yan Zhang, and Hui Li. Mmict: Boosting multi-modal fine-tuning with in-context examples, 2023.
- [58] Jihao Qiu, Yuan Zhang, Xi Tang, Lingxi Xie, Tianren Ma, Pengyu Yan, David Doermann, Qixiang Ye, and Yunjie Tian. Artemis: Towards referential understanding in complex videos. *arXiv preprint arXiv:2406.00258*, 2024.
- [59] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2024.
- [60] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorf: Incentivizing video reasoning capability in mllms via reinforced fine-tuning, 2025.
- [61] Yunlong Tang, Gen Zhan, Li Yang, Yiting Liao, and Chenliang Xu. Cardiff: Video salient object ranking chain of thought reasoning for saliency prediction with diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 7302–7310, 2025.
- [62] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*, 2024.
- [63] Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. *arXiv preprint arXiv:2506.08817*, 2025.
- [64] Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning, 2025.
- [65] Ruizhe Chen, Zhiting Fan, Tianze Luo, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video temporal grounding via reinforcement learning. *arXiv preprint arXiv:2507.18100*, 2025.

- [66] Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, and Weiyao Lin. Mecd: Unlocking multi-event causal discovery in video reasoning, 2024.
- [67] Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025.
- [68] Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding, 2025.
- [69] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding, 2025.
- [70] Jiaao Li, Kaiyuan Li, Chen Gao, Yong Li, and Xinlei Chen. Egoprune: Efficient token pruning for egomotion video reasoning in embodied agent, 2025.
- [71] Ashwin Vinod, Shrey Pandit, Aditya Vavre, and Linshen Liu. Egovlm: Policy optimization for egocentric video understanding, 2025.
- [72] Mingxian Lin, Wei Huang, Yitang Li, Chengjie Jiang, Kui Wu, Fangwei Zhong, Shengju Qian, Xin Wang, and Xiaojuan Qi. Embrace-3k: Embodied reasoning and action in complex environments, 2025.
- [73] Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos, 2024.
- [74] Jiaxin Liu and Zhaolu Kang. Reasonact: Progressive training for fine-grained video reasoning in small models, 2025.
- [75] Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. St-think: How multimodal large language models reason about 4d worlds from ego-centric videos, 2025.
- [76] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, Shixiong Zhao, Shuai Peng, Shuangye Li, Sihang Yuan, Sijin Wu, Tianheng Cheng, Weiwei Liu, Wenqian Wang, Xianhan Zeng, Xiao Liu, Xiaobo Qin, Xiaohan Ding, Xiaojun Xiao, Xiaoying Zhang, Xuanwei Zhang, Xuehan Xiong, Yanghua Peng, Yangrui Chen, Yanwei Li, Yanxu Hu, Yi Lin, Yiyuan Hu, Yiyuan Zhang, Youbin Wu, Yu Li, Yudong Liu, Yue Ling, Yujia Qin, Zanbo Wang, Zhiwu He, Aoxue Zhang, Bairen Yi, Bencheng Liao, Can Huang, Can Zhang, Chaorui Deng, Chaoyi Deng, Cheng Lin, Cheng Yuan, Chenggang Li, Chenhui Gou, Chenwei Lou, Chengzhi Wei, Chundian Liu, Chunyuan Li, Deyao Zhu, Donghong Zhong, Feng Li, Feng Zhang, Gang Wu, Guodong Li, Guohong Xiao, Haibin Lin, Haihua Yang, Haoming Wang, Heng Ji, Hongxiang Hao, Hui Shen, Huixia Li, Jiahao Li, Jialong Wu, Jianhua Zhu, Jianpeng Jiao, Jiashi Feng, Jiaze Chen, Jianhui Duan, Jihao Liu, Jin Zeng, Jingqun Tang, Jingyu Sun, Joya Chen, Jun Long, Junda Feng, Junfeng Zhan, Junjie Fang, Junting Lu, Kai Hua, Kai Liu, Kai Shen, Kaiyuan Zhang, Ke Shen, Ke Wang, Keyu Pan, Kun Zhang, Kunchang Li, Lanxin Li, Lei Li, Lei Shi, Li Han, Liang Xiang, Liangqiang Chen, Lin Chen, Lin Li, Lin Yan, Liying Chi, Longxiang Liu, Mengfei Du, Mingxuan Wang, Ningxin Pan, Peibin Chen, Pengfei Chen, Pengfei Wu, Qingqing Yuan, Qingyao Shuai, Qiuyan Tao, Renjie Zheng, Renrui Zhang, Ru Zhang, Rui Wang, Rui Yang, Rui Zhao, Shaoqiang Xu, Shihao Liang, Shipeng Yan, Shu Zhong, Shuaishuai Cao, Shuangzhi Wu, Shufan Liu, Shuhan Chang,

Songhua Cai, Tenglong Ao, Tianhao Yang, Tingting Zhang, Wanjun Zhong, Wei Jia, Wei Weng, Weihao Yu, Wenhao Huang, Wenjia Zhu, Wenli Yang, Wenzhi Wang, Xiang Long, XiangRui Yin, Xiao Li, Xiaolei Zhu, Xiaoying Jia, Xijin Zhang, Xin Liu, Xinchen Zhang, Xinyu Yang, Xiongcai Luo, Xiuli Chen, Xuantong Zhong, Xuefeng Xiao, Xujing Li, Yan Wu, Yawei Wen, Yifan Du, Yihao Zhang, Yining Ye, Yonghui Wu, Yu Liu, Yu Yue, Yufeng Zhou, Yufeng Yuan, Yuhang Xu, Yuhong Yang, Yun Zhang, Yunhao Fang, Yuntao Li, Yurui Ren, Yuwen Xiong, Zehua Hong, Zehua Wang, Zewei Sun, Zeyu Wang, Zhao Cai, Zhaoyue Zha, Zhecheng An, Zhehui Zhao, Zhengzhuo Xu, Zhipeng Chen, Zhiyong Wu, Zhuofan Zheng, Zihao Wang, Zilong Huang, Ziyu Zhu, and Zuquan Song. Seed1.5-vl technical report, 2025.

- [77] Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. *arXiv preprint arXiv:2504.13122*, 2025.
- [78] Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding, 2025.
- [79] Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun MA, and Chao Zhang. video-salmonn-o1: Reasoning-enhanced audio-visual large language model, 2025.
- [80] Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation, 2025.
- [81] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J. Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo, 2025.
- [82] Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking, 2025.
- [83] Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via r1-zero-like training, 2025.
- [84] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning, 2025.
- [85] Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, and Limin Wang. Videocap-r1: Enhancing mllms for video captioning via structured thinking, 2025.
- [86] Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency, 2025.
- [87] Fanrui Zhang, Dian Li, Qiang Zhang, Chenjun, sinbadliu, Junxiong Lin, Jiahong Yan, Jiawei Liu, and Zheng-Jun Zha. Fact-r1: Towards explainable video misinformation detection with deep reasoning, 2025.
- [88] Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning, 2025.
- [89] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning, 2025.

- [90] Siran Chen, Boyu Chen, Chenyun Yu, Yuxiao Luo, Ouyang Yi, Lei Cheng, Chengxiang Zhuo, Zang Li, and Yali Wang. Vragent-r1: Boosting video recommendation with mllm-based agents via reinforcement learning, 2025.
- [91] Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning, 2025.
- [92] Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms, 2025.
- [93] Yifeng Gao, Yifan Ding, Hongyu Su, Juncheng Li, Yunhan Zhao, Lin Luo, Zixing Chen, Li Wang, Xin Wang, Yixu Wang, Xingjun Ma, and Yu-Gang Jiang. David-xr1: Detecting ai-generated videos with explainable reasoning, 2025.
- [94] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025.
- [95] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. *arXiv preprint arXiv:2507.15028*, 2025.
- [96] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding, 2025.
- [97] Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames, 2025.
- [98] Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Bohan Zeng, Yang Shi, Sihan Yang, Pengfei Wan, Qiang Liu, Liang Wang, and Tieniu Tan. Versavid-r1: A versatile video understanding and reasoning model from question answering to captioning tasks, 2025.
- [99] Tony Montes and Fernando Lozano. Viqagent: Zero-shot video question answering via agent with open-vocabulary grounding validation, 2025.
- [100] Yiran Meng, Junhong Ye, Wei Zhou, Guanghui Yue, Xudong Mao, Ruomei Wang, and Baoquan Zhao. Videoforest: Person-anchored hierarchical reasoning for cross-video question answering, 2025.
- [101] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning, 2025.
- [102] Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. Video-skill-cot: Skill-based chain-of-thoughts for domain-adaptive video reasoning, 2025.
- [103] Sunqi Fan, Meng-Hao Guo, and Shuojin Yang. Agentic keyframe search for video question answering, 2025.
- [104] Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. Wait, we don't need to "wait"! removing thinking tokens improves reasoning efficiency, 2025.

- [105] Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. Vca: Video curious agent for long video understanding, 2025.
- [106] Ji Soo Lee, Jongha Kim, Jeehye Na, Jinyoung Park, and Hyunwoo J. Kim. Vidchain: Chain-of-tasks with metric-based direct preference optimization for dense video captioning, 2025.
- [107] Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Ji-Rong Wen, and Zhicheng Dou. Videodeepresearch: Long video understanding with agentic tool using, 2025.
- [108] Jiahao Meng, Shuyang Sun, Yue Tan, Lu Qi, Yunhai Tong, Xiangtai Li, and Longyin Wen. Cyberv: Cybernetics for test-time scaling in video understanding, 2025.
- [109] Kangsan Kim, Geon Park, Youngwan Lee, Woongyeong Yeo, and Sung Ju Hwang. Videoicl: Confidence-based iterative in-context learning for out-of-distribution video understanding, 2024.
- [110] Umihiro Kamoto, Tatsuya Ishibashi, and Noriyuki Kugo. Dive: Deep-search iterative video exploration a technical report for the cvrr challenge at cvpr 2025, 2025.
- [111] Linhao Yu, Xinguang Ji, Yahui Liu, Fanheng Kong, Chenxi Sun, Jingyuan Zhang, Hongzhi Zhang, V. W. Wang, Fuzheng Zhang, and Deyi Xiong. Evaluating multimodal large language models on video captioning via monte carlo tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. AutoCaption + MCTS and the MCTS-VCB benchmark.
- [112] Ziqi Pang and Yu-Xiong Wang. Mr. video: "mapreduce" is the principle for long video understanding, 2025.
- [113] Kuo Wang, Quanlong Zheng, Junlin Xie, Yanhao Zhang, Jinguo Luo, Haonan Lu, Liang Lin, Fan Zhou, and Guanbin Li. Free-moref: Instantly multiplexing context perception capabilities of video-mllms within single inference, 2025.
- [114] Ce Zhang, Yan-Bo Lin, Ziyang Wang, Mohit Bansal, and Gedas Bertasius. Silvr: A simple language-based video reasoning framework, 2025.
- [115] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhui Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning, 2025.
- [116] Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. Reagent-v: A reward-driven multi-agent framework for video understanding, 2025.
- [117] Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Ji-Rong Wen, and Zhicheng Dou. Videodeepresearch: Long video understanding with agentic tool using. *arXiv preprint arXiv:2506.10821*, 2025.
- [118] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding, 2025.
- [119] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

- [120] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.
- [121] Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.
- [122] Xinhao Li, Zhenpeng Huang, Jing Wang, Kunchang Li, and Limin Wang. Videoeval: Comprehensive benchmark suite for low-cost evaluation of video foundation model. *arXiv preprint arXiv:2407.06491*, 2024.
- [123] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024.
- [124] Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, Pooyan Fazli, and Chenliang Xu. Vidcomposition: Can mllms analyze compositions in compiled videos? In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8490–8500, 2025.
- [125] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning, 2025.
- [126] Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y. Charles, Xinyu Zhou, and Xu Sun. Videoreasonbench: Can mllms perform vision-centric complex video reasoning?, 2025.
- [127] Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, Cordelia Schmid, and Tobias Weyand. Minerva: Evaluating complex video reasoning, 2025.
- [128] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding, 2025.
- [129] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.
- [130] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019.
- [131] Heqing Zou, Tianze Luo, Guiyang Xie, Victor Xiao Jie Zhang, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, and Huaijian Zhang. Hlv-1k: A large-scale hour-long video benchmark for time-specific long video understanding, 2025.
- [132] David Ma, Huaqing Yuan, Xingjian Wang, Qianbo Zang, Tianci Liu, Xinyang He, Yanbin Wei, Jiawei Guo, Ni Jiahui, Zhenzhu Yang, Meng Cao, Shanghaoran Quan, Yizhi Li, Wangchunshu Zhou, Jiaheng Liu, Wenhao Huang, Ge Zhang, Shiwen Ni, and Xiaojie Jin. Scalelong: A multi-timescale benchmark for long video understanding, 2025.

- [133] Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding, 2025.
- [134] Zicheng Zhao, Kangyu Wang, Shijie Li, Rui Qian, Weiyao Lin, and Huabin Liu. Cogstream: Context-guided streaming video question answering, 2025.
- [135] Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts, 2025.
- [136] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [137] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multi-modal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023.
- [138] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023.
- [139] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.
- [140] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023.
- [141] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023.
- [142] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens, 2023.
- [143] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering, 2024.
- [144] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023.
- [145] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.
- [146] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvilm: Efficient long video understanding via large language models. *arXiv preprint arXiv:2404.03384*, 2024.
- [147] Yizhou Wang, Ruiyi Zhang, Haoliang Wang, Uttaran Bhattacharya, Yun Fu, and Gang Wu. Vaquita: Enhancing alignment in llm-assisted video understanding, 2023.

- [148] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.
- [149] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-lmms. *arXiv preprint arXiv:2406.07476*, 2024.
- [150] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.
- [151] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univ: Unified visual representation empowers large language models with image and video understanding, 2024.
- [152] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024.
- [153] Honglu Zhou, Xiangyu Peng, Shrikant Kendre, Michael S. Ryoo, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. Strefer: Empowering video lmms with space-time referring and reasoning via synthetic instruction data, 2025.
- [154] Yunxiao Wang, Meng Liu, Wenqi Liu, Xuemeng Song, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, Guorui Zhou, and Liqiang Nie. Time: Temporal-sensitive multi-dimensional instruction tuning and robust benchmarking for video-lmms, 2025.
- [155] Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with lmms, 2024.
- [156] Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models, 2024.
- [157] Jen-Hao Cheng, Vivian Wang, Huayu Wang, Huapeng Zhou, Yi-Hao Peng, Hou-I Liu, Hsiang-Wei Huang, Kuang-Ming Chen, Cheng-Yen Yang, Wenhao Chai, Yi-Ling Chen, Vibhav Vineet, Qin Cai, and Jenq-Neng Hwang. Tempura: Temporal event masked prediction and understanding for reasoning in action, 2025.
- [158] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.
- [159] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024.
- [160] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [161] Qiji Zhou, Yifan Gong, Guangsheng Bao, Hongjie Qiu, Jinqiang Li, Xiangrong Zhu, Huajian Zhang, and Yue Zhang. Reasoning is all you need for video generalization: A counterfactual benchmark with sub-question evaluation, 2025.

- [162] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning, 2025.
- [163] Chao Huang, Benfeng Wang, Jie Wen, Chengliang Liu, Wei Wang, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought, 2025.
- [164] Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. Compile scene graphs with reinforcement learning, 2025.
- [165] Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Busterx: Mllm-powered ai-generated video forgery detection and explanation, 2025.
- [166] Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, Jinwen Luo, Weibo Gu, Zexuan Li, Xiaojing Zhang, Yangyu Tao, Han Hu, Di Wang, and Ying Shan. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts, 2025.
- [167] Hosu Lee, Junho Kim, Hyunjun Kim, and Yong Man Ro. Refocus: Reinforcement-guided frame optimization for contextual understanding, 2025.
- [168] Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, Fan Yang, Guorui Zhou, Hao Peng, Haojie Ding, Jiaming Huang, Jiangxia Cao, Jiankang Chen, Jingyun Hua, Jin Ouyang, Kaibing Chen, Kaiyu Jiang, Kaiyu Tang, Kun Gai, Shengnan Zhang, Siyang Mao, Sui Huang, Tianke Zhang, Tingting Gao, Wei Chen, Wei Yuan, Xiangyu Wu, Xiao Hu, Xingyu Lu, Yang Zhou, Yi-Fan Zhang, Yiping Yang, Yulong Chen, Zhenhua Wu, Zhenyu Li, Zhixin Ling, Ziming Li, Dehua Ma, Di Xu, Haixuan Gao, Hang Li, Jiawei Guo, Jing Wang, Lejian Ren, Muhaow Wei, Qianqian Wang, Qigen Hu, Shiyao Wang, Tao Yu, Xinchen Luo, Yan Li, Yiming Liang, Yuhang Hu, Zeyi Lu, Zhuoran Yang, and Zixing Zhang. Kwai keye-vl technical report, 2025.
- [169] Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration, 2025.
- [170] Mengjingcheng Mo, Xinyang Tong, Jiaxu Leng, Mingpi Tan, Jiankang Zheng, Yiran Liu, Haosheng Chen, Ji Gan, Weisheng Li, and Xinbo Gao. A2seek: Towards reasoning-centric benchmark for aerial anomaly understanding, 2025.
- [171] Feng Yue, Zhaoxing Zhang, Junming Jiao, Zhengyu Liang, Shiwen Cao, Feifei Zhang, and Rong Shen. Tempo-r0: A video-mllm for temporal video grounding through efficient temporal sensing reinforcement learning, 2025.
- [172] Yiwei Sun, Peiqi Jiang, Chuanbin Liu, Luohao Lin, Zhiying Lu, and Hongtao Xie. From evaluation to defense: Advancing safety in video large language models, 2025.
- [173] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.
- [174] Yogesh Kulkarni and Pooyan Fazli. Avatar: Reinforcement learning to see, hear, and reason over video. *arXiv preprint arXiv:2508.03100*, 2025.

- [175] Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. Framemind: Frame-interleaved chain-of-thought for video reasoning via reinforcement learning. *arXiv preprint arXiv:2509.24008*, 2025.
- [176] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025.
- [177] Xiao Wang, Liye Jin, Xufeng Lou, Shiao Wang, Lan Chen, Bo Jiang, and Zhipeng Zhang. Reasoningtrack: Chain-of-thought reasoning for long-term vision-language tracking. *arXiv preprint arXiv:2508.05221*, 2025.
- [178] Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*, 2025.
- [179] Sicheng Tao, Jungang Li, Yibo Yan, Junyan Zhang, Yubo Gao, Hanqian Li, ShuHang Xun, Yuxuan Fan, Hong Chen, Jianxiang He, et al. Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning. *arXiv preprint arXiv:2509.21113*, 2025.
- [180] Chaohong Guo, Xun Mo, Yongwei Nie, Xuemiao Xu, Chao Xu, Fei Yu, and Chengjiang Long. Tar-tvg: Enhancing vlms with timestamp anchor-constrained reasoning for temporal video grounding. *arXiv preprint arXiv:2508.07683*, 2025.
- [181] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025.
- [182] Yiwei Sun, Peiqi Jiang, Chuanbin Liu, Luohao Lin, Zhiying Lu, and Hongtao Xie. From evaluation to defense: Advancing safety in video large language models. *arXiv preprint arXiv:2505.16643*, 2025.
- [183] Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning. *arXiv preprint arXiv:2506.00318*, 2025.
- [184] Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*, 2025.
- [185] Xu Yang, Qi Zhang, Shuming Jiang, Yaowen Xu, Zhaofan Zou, Hao Sun, and Xuelong Li. Meter: Multi-modal evidence-based thinking and explainable reasoning–algorithm and benchmark. *arXiv preprint arXiv:2507.16206*, 2025.
- [186] Sitong Gong, Lu Zhang, Yunzhi Zhuge, Xu Jia, Pingping Zhang, and Huchuan Lu. Reinforcing video reasoning segmentation to think before it segments. *arXiv preprint arXiv:2508.11538*, 2025.
- [187] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025.
- [188] Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. Adsqa: Towards advertisement video understanding. *arXiv preprint arXiv:2509.08621*, 2025.

- [189] Kehua Chen. Chronoforge-rl: Chronological forging through reinforcement learning for enhanced video understanding. *arXiv preprint arXiv:2509.15800*, 2025.
- [190] Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms. *arXiv preprint arXiv:2509.18056*, 2025.
- [191] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [192] Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. Streamvln: Streaming vision-and-language navigation via slowfast context modeling, 2025.
- [193] Ali Vosoughi, Jing Bi, Pinxin Liu, Yunlong Tang, and Chenliang Xu. Can sound replace vision in llava with token substitution? *arXiv preprint arXiv:2506.10416*, 2025.
- [194] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning?, 2025.
- [195] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.
- [196] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023.
- [197] Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. Ovo-bench: How far is your video-llms from real-world online video understanding?, 2025.
- [198] Ruijie Wu, Xiaojian Ma, Hai Ci, Yue Fan, Yuxuan Wang, Haozhe Zhao, Qing Li, and Yizhou Wang. Longvitu: Instruction tuning for long-form video understanding, 2025.
- [199] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024.
- [200] Zhoufaran Yang, Yan Shu, Zhifei Yang, Yan Zhang, Yu Li, Keyang Lu, Gangyan Zeng, Shaohui Liu, Yu Zhou, and Nicu Sebe. Vidtext: Towards comprehensive evaluation for video text understanding, 2025.
- [201] Chenkai Zhang, Yiming Lei, Zeming Liu, Haitao Leng, Shaoguo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. Seriesbench: A benchmark for narrative-driven drama series understanding, 2025.
- [202] L'ea Dubois, Klaus Schmidt, Chengyu Wang, Ji-Hoon Park, Lin Wang, and Santiago Munoz. Video event reasoning and prediction by fusing world knowledge from llms with vision foundation models, 2025.

- [203] Enxin Song, Wenhao Chai, Shusheng Yang, Ethan Armand, Xiaojun Shan, Haiyang Xu, Jianwen Xie, and Zhuowen Tu. Videonsa: Native sparse attention scales video understanding. *arXiv preprint arXiv:2510.02295*, 2025.
- [204] Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, YX Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*, 2025.
- [205] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [206] Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhui Chen. Videoeval-pro: Robust and realistic long video understanding evaluation, 2025.
- [207] Yogesh Kulkarni and Pooyan Fazli. Videosavi: Self-aligned video language models without human supervision, 2025.
- [208] Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning, 2025.
- [209] Ziyi Wang, Haoran Wu, Yiming Rong, Deyang Jiang, Yixin Zhang, Yunlong Zhao, Shuang Xu, and Bo XU. Lvc: A lightweight compression framework for enhancing vlms in long video understanding, 2025.