Eagle 2.5: Boosting Long-Context Post-Training for Frontier Vision-Language Models

Guo Chen[1,2*], **Zhiqi Li**[1,2], **Shihao Wang**[3,2], **Jindong Jiang**[2], **Yicheng Liu**[1], **Lidong Lu**[1], **De-An Huang**[2], **Wonmin Byeon**[2], **Matthieu Le**[2], **Max Ehrlich**[2], **Tong Lu**[1†], **Limin Wang**[1†], **Bryan Catanzaro**[2], **Jan Kautz**[2], **Andrew Tao**[2], **Zhiding Yu**[2†], **Guilin Liu**[2†]
[1]Nanjing University, [2]NVIDIA, [3]Hong Kong Polytechnic University
chenguo1177@gmail.com

Abstract

We introduce Eagle2.5, a frontier vision-language model (VLM) for long-context multimodal learning. Our work addresses the challenges in long video comprehension and high-resolution image understanding, introducing a generalist framework for both tasks. The proposed training framework incorporates Automatic Degrade Sampling and Image Area Preservation, two techniques that preserve contextual integrity and visual details. The framework also includes numerous efficiency optimizations in the pipeline for long-context data training. Finally, we propose Eagle-Video-110K, a novel dataset that integrates both story-level and clip-level annotations, facilitating long-video understanding. Eagle2.5 demonstrates substantial improvements on long-context multimodal benchmarks, providing a robust solution to the limitations of existing VLMs. Notably, our best model Eagle2.5-8B achieves 72.4% on Video-MME with 512 input frames, matching the results of top-tier commercial model such as GPT-4o and large-scale open-source models like Qwen2.5-VL-72B and InternVL2.5-78B.

1 Introduction

Despite the significant advances in multimodal learning [4, 16, 52, 53, 97], many vision-language models (VLMs) remain focused on short-context tasks, with long-context understanding under-explored. This gap is particularly evident in both long video comprehension and high-resolution image/video understanding, where the processing of extended visual contexts remains an open challenge. Such extended contexts encompass multiple images, extended video sequences, high-resolution media, or combinations thereof. However, the development of long-context VLMs is still in its early stages, hindered by fundamental challenges in dataset construction, architecture design, training strategies, and computation/memory bottlenecks.

To enable long-context visual understanding, several approaches have been proposed to address the challenge of processing extended visual inputs by designing specialized compression or selection modules [52, 84, 57, 45, 43, 117, 103, 53]. While these methods effectively circumvent the need to extend the context length of VLMs, they often introduce additional computational overhead or capacity limitations, potentially constraining model performance. A promising research direction is to extend the context length of LLMs to enable native long-context understanding. While prior studies [110, 126, 85] have explored this direction, challenges and key limitations still remain. First, the performance of existing methods is often suboptimal, generally falling behind proprietary models. Second, these approaches struggle to achieve consistent improvements as the amount of visual input

[*]Work done during an internship at NVIDIA.
[†]Co-corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

increases. Lastly, the optimal training strategies for state-of-the-art long-context VLMs remain unclear, given the complex interplay of factors such as training strategies and data recipes.

To this end, we present Eagle 2.5, a versatile multimodal model designed to efficiently process extensive contextual information. Unlike models solely optimized for handling long multimodal sequences without improving performance, Eagle-2.5 benefits from increased input length, leading to consistent performance gains besides merely accommodating longer inputs. As shown in Fig. 1, our model achieves superior context coverage and exhibits consistent performance scaling with increasing frame counts. Notably, it attains competitive results compared to larger models such as GPT-4o [77] and Qwen2.5-VL-72B [5], while maintaining a significantly smaller parameter footprint.

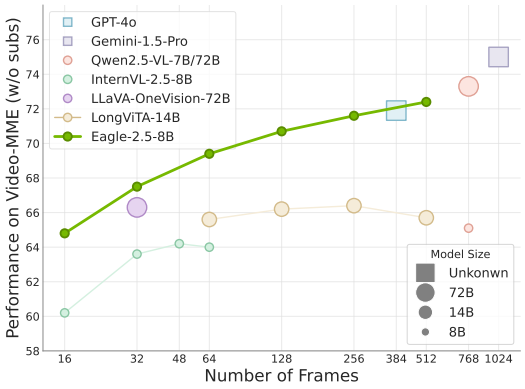

Figure 1: **Performance comparison of Eagle2.5 with leading vision-language models on the Video-MME benchmark.** Eagle2.5 demonstrates consistent improvement as the number of input frames increases.

Eagle 2.5 is driven by both the advanced training strategy and data recipe. For training strategy, we introduce two core components for effective long-context learning: information-first sampling and progressive training.

- **Information-first sampling**. The information-first sampling strategy ensures the preservation of essential visual and semantic information through two mechanisms: (1) *Image Area Preservation*, which optimizes tiling to retain the majority of the original image area while maintaining aspect ratio fidelity, avoiding rigid aspect ratio constraints; and (2) *Automatic Degradation Sampling (ADS)*, which dynamically balances visual and textual inputs by prioritizing complete text retention while adaptively optimizing visual content to maximize context length utilization and preserve multimodal information.
- **Progressive training**. We employ a progressive mixed post-training approach, wherein context length is incrementally expanded during training, enhancing the model's ability to process inputs of varying sizes. This integrated strategy significantly improves information density over static sampling methods while ensuring consistent performance across diverse input types and lengths.

For data recipe, we embrace the "*diversity first, then quality*" principle in curating the training data pool. Our data recipe combines open-source data (including human-annotated data as well as synthetic video data) with our self-curated Eagle-Video-110K dataset, specifically designed to enhance long video understanding capabilities. We adopt a diversity-driven collection strategy, using multiple video sources and a similarity thresholding method to identify novel clips that maximize content diversity. Our dataset is distinguished by its dual annotation approach:

- A top-down story-level method that leverages human-annotated chapters as meaningful segments instead of traditional shot-level segmentation, producing dense captions that form the basis for comprehensive long-form QA pairs capturing the entire video's narrative structure;
- A complementary bottom-up clip-level approach that generates focused QA pairs for short clips using GPT-4o with diverse question types. To address the challenge of extending localized clip annotations to full-length videos, we implement anchors that incorporates temporal references and contextual elements without revealing answers, thereby letting models understand both overarching narratives and precise spatio-temporal details within videos.

2 Related Work

Vision-language models. Advancements in large language models (LLMs) [20, 1, 77] have significantly propelled visual understanding by integrating visual features, leading to the creation of Visual Language Models (VLMs) [52, 76, 58, 133]. Open-source VLMs [58, 60, 49, 97, 92, 18, 66, 50, 114, 67, 39] continue to achieve breakthroughs, often matching or exceeding the performance of state-of-the-art commercial models like GPT-4V/4o [77] and Gemini-1.5 [83]. The release of open-source VLMs [49, 95, 40, 38], complete with its training data and code base, has further

accelerated research in this area. However, most current VLMs primarily focus on short-context understanding, handling only a few images or short video clips at a time. Eagle 2.5 advances this field by concentrating on long-context visual understanding through a comprehensive exploration and development of training strategies and data recipes.

Long-context VLMs. Long-context VLMs were developed to address the challenges of processing large multimodal sequences. Currently, methods for long-context VLMs fall into two main categories. The first category involves specialized modules designed for context compression. Question-guided compressions [84, 57, 45, 98] or selection [117, 55, 116] methods extract question-related visual cues through an additional module, while various token reduction techniques [43, 117, 103, 53, 57, 124, 102, 62, 62, 61] aim to minimize the visual representation before LLM processing. The other category attempts to directly extend the context of LLMs. Works like LongVA [126], LongVILA [110], and LongViTA [85] extend the context length of LLMs to accommodate longer multimodal sequences. While promising, these approaches often underperform proprietary models, fail to show consistent performance improvements with increasing visual input, and have underexplored constraints on training strategies and data recipes. Our approach focuses on developing native long-context capabilities that enhance VLMs by exploring training data, formulations, and without introducing additional compression modules or suffering from performance inconsistencies observed in previous expansion attempts.

Long-context multimodal data. To enhance VLMs' long-context multimodal understanding capabilities, various datasets have been proposed. Some datasets focus on multimodal understanding of long documents [89, 96, 94, 80], such as slides and papers. However, they often lack temporal understanding. Other datasets [36, 105, 24, 82, 87, 88, 120, 121] emphasize the temporal coherence and information retrieval across long spans inherent in movies. Additionally, recent datasets [73, 13, 30, 129] covering domains further enhance VLMs' long-context multimodal understanding. Regarding the annotation methods for long-context multimodal datasets, early works [89, 96, 94, 36, 87, 73] relied on manual efforts. To reduce costs, some methods [80, 24, 82, 88, 120, 121, 30, 13, 129] use tools like GPT-4V [76] and Gemini [91] for automated or semi-automated annotation. Recent advancements in data construction emphasize hierarchical annotation strategies [30], which can preserve narrative structure in long videos. These advancements reflect a trend towards creating balanced datasets that effectively assess long-context multimodal understanding while managing creation costs.

3 Method

This section introduces the model architecture, training strategies, and data recipe of Eagle2.5.

3.1 Model Architecture

We design our proposed model as a versatile multimodal system capable of efficiently processing long-context information, rather than a specialized model solely optimized for handling extended multimodal inputs. To ensure adaptability and generalization across diverse tasks, we deliberately avoid incorporating tailored compression modules that might constrain the model's flexibility. Following the architecture of LLaVA [58], we employ an MLP projection layer to align vision embeddings from SigLIP [123] with the LLM representation space, as shown in Fig. 2. In this work, we utilize the Qwen2.5 series models [93]. To effectively handle any-resolution images, we adopt the image tiling strategy, inspired by LLaVA-1.5 [59] and InternVL [16].

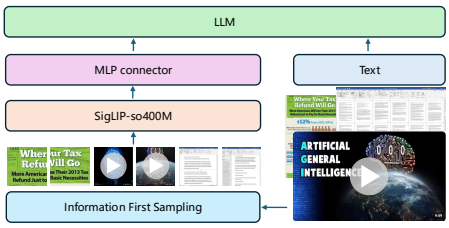

Figure 2: Tiling-based general multimodal system.

3.2 Training Strategy

Our approach contains two key components to achieve effective long-context training: first, an information-first sampling strategy that establishes optimal sampling criteria; and second, a progressive training schedule based on this strategy, which directs the entire model training process.

3.2.1 Information-First Sampling

In multimodal training, the sampling of visual content is essential. Multi-image documents typically comprise dozens of pages with ultra-high-resolution images, while video content can vary drastically in length - from mere seconds to hours. To effectively manage this diversity, we present **information-first sampling** to promote information preservation from both visual and semantic dimensions.

Image area preservation (IAP). Traditional tiling methods divide an image of size $W \times H$ into a rigid $r_w \times r_h$ grid of $s \times s$ tiles. While effective for handling high-resolution inputs, these approaches often distort the original image geometry through improper aspect ratio handling. For example, InternVL [16] imposes strict aspect ratio constraints that force image downsampling, undermining the purpose of tiling. To address this, we propose an area-prioritized tiling strategy that optimizes two key objectives:

- *Area Preservation*: Encourage maintaining at least 60% of the original area ($A_{\text{orig}} = WH$) in the tiled version ($A_{\text{new}} = r_w r_h s^2$).
- *Aspect Ratio Fidelity*: Align the tiling ratio $r_t = r_w / r_h$ with the original aspect ratio $r_{\text{orig}} = W/H$.

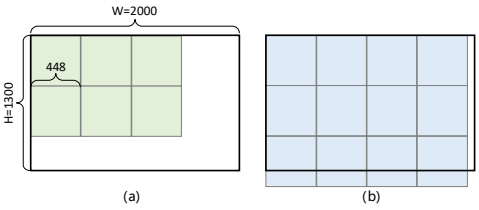

Figure 3: **Image area preservation.** Compared to the tiling strategy (a) from InternVL [92], our method (b) effectively retains a larger portion of the original image, especially for high-resolution inputs. This ensures that more comprehensive visual information is preserved, benefiting tasks that require fine-grained details.

For candidate tiling ratios $\{(r_w, r_h) \mid r_w \times r_h \leq N\}$, we select the optimal configuration by:

$$\arg\max_{(r_w, r_h)} \left[\underbrace{\min\left(\frac{A_{\text{new}}}{A_{\text{orig}}}, 0.6\right)}_{\text{Area penalty}} \cdot \underbrace{\min\left(\frac{r_t}{r_{\text{orig}}}, \frac{r_{\text{orig}}}{r_t}\right)}_{\text{Aspect ratio alignment}} \right] \tag{1}$$

This formulation imposes penalties when $A_{\text{new}} < 0.6 A_{\text{orig}}$ but avoids over-rewarding configurations where $A_{\text{new}} > 0.6 \times A_{\text{orig}}$. The aspect ratio term reaches maximum value 1 when $r_t = r_{\text{orig}}$, decaying symmetrically for deviations. A comparison between the strategies is shown in Fig. 3.

Automatic degradation sampling. VLMs require careful allocation of sequence length budgets between visual and textual inputs. Conventional **vision-context-centric** approaches sample visual content (e.g., video frames) at fixed rates or with predetermined counts, risking text truncation and suboptimal token allocation. We propose Automatic Degradation Sampling (**ADS**), an **all-context-centric** strategy that dynamically optimizes this balance.

Given a training sample $\mathcal{S} = \{S_{\text{visual}}, S_{\text{text}}\}$ with max sequence length $\mathcal{L}_{\max}$, where S_{visual} contains arbitrary combinations of images, videos, and multi-page documents: 1) Compute fixed text token length $\mathcal{L}_{\text{text}}$; and Derive fixed visual token budget: $\mathcal{L}_{\text{visual}} = \mathcal{L}_{\max} - \mathcal{L}_{\text{text}}$. Thus, we keep the complete textual information by constricting the visual token budget.

For visual content optimization under $\mathcal{L}_{\text{visual}}$, we distinguish two types and optimize two key variables:

- **Images**: Optimize *maximal tile count per image* t to maximize spatial information of M images.
- **Temporal content** (video/doc): Optimize *sampling count* n to maximize temporal coverage.

The constrained optimization problem is formulated as:

$$\begin{aligned} &\max_{1 \leq t \leq 12,\ 1 \leq n \leq N_{\max}} \sum_{i=1}^{M} L(t, I_i) + 256n \\ &\text{s.t.} \sum_{i=1}^{M} L(t, I_i) + 256n \leq \mathcal{L}_{\text{vis}} \end{aligned} \tag{2}$$

Where **optimization variables** are the tile count per image (t) and temporal sampling count (n), with **fixed parameters** including: total image instances M (calculated from input), token function $L(t, I_i)$

used to calculate the tokens of i-th image I_i under maximal tiling number t, predefined upper bounds $T_{\max} = 12$ (max tiles per image) and $N_{\max} = 2 \times \text{duration}/1 \times \text{pages}$ (video/doc constraints). For temporal content, we do not use image tiling, thus the token quantity per temporal unit (frame/page) is $L(1, \cdot) = 256$.

Given that training samples typically exhibit mutually exclusive composition (predominantly images *or* temporal content), ADS employs a dual-phase degradation process to address the above optimization problem:

- *Temporal degradation*: Initially, we fix the max tile number $t = 1$ and focus on temporal sampling. We target a sampling rate of 2 FPS for videos, and the usage of all images for multi-image documents. We also require that each visual input has at least $N_{\min}$ frames; if this minimum cannot be met within the visual context budget, the sample is discarded. Formally, the maximally sampled temporal units is $n^* = \left\lfloor \frac{\mathcal{L}_{\text{visual}} - M \times 256}{256} \right\rfloor$.
- *Tiling degradation*: After deciding the number of frames, we dynamically adjust the tiling to maximize the use of available context. Let $\mathcal{T} = \{12, 8, 6, 4, 2, 1\}$ represent the possible tile configurations in decreasing order. We choose the highest tile configuration t^* such that: $t^* = \max \{t \in \mathcal{T} : \sum_{i=1}^{m} L(t, I_i) \leq (\mathcal{L}_V - n^* \times 256)\}$ This strategy preserves as much visual detail as possible while ensuring the full textual input is retained, thereby optimizing the overall learning signal.

This dual-phase approach guarantees complete text preservation while dynamically adapting visual resolution to available context budget, achieving superior information density compared to static sampling strategies.

3.2.2 Post-Training Schedule

We introduce a comprehensive post-training framework consisting of two complementary strategies. First, we establish a foundational mixed post-training approach, upon which we develop an enhanced progressive mixed post-training strategy to substantially improve model performance across varying context lengths.

- *Mixed post-training.* Since the model needs to efficiently process multimodal inputs of diverse lengths, maintaining consistent performance across variable context sizes is essential. Our ADS method adaptively adjusts each training sample to the maximum sequence length $\mathcal{L}_{\text{max}}$, providing a frame-agnostic training paradigm. We implement a mixed training strategy with length-balanced packing [3] to optimize performance uniformly across the entire spectrum of context lengths.
- *Progressive mixed post-training.* For scenarios with large $\mathcal{L}_{\text{max}}$ values, balancing the distribution of long and short sequences becomes computationally intensive, and achieving optimal performance through a single training iteration proves challenging. To address this limitation, we propose a progressive mixed training methodology that gradually exposes the model to increasingly larger $\mathcal{L}_{\text{max}}$ values, systematically enhancing its capacity to process extended contexts. Compared to conventional mixed training, our method more effectively preserves the model's capabilities across different sequence lengths while safely generating diverse model variants at intermediate training stages. In our exeriment, we sequentially set $\mathcal{L}_{\text{max}}$ to 32K, 64K and 128K.

3.3 Data Recipe

Our data recipe begins with open-source data. We embrace the "diversity first, then quality" principle and gather data from various open sources. This data mainly comprises high-definition multi-image/short videos, long videos, multi-page documents, and extensive text data. We also find that current open-source video data often lacks sufficient length. We thus propose a novel dataset, Eagle-Video-110K, to complement the length, as shown in Fig. 4.

3.3.1 Open-Source Long-Context Data

A model's capability is intrinsically linked to the diversity of its training data. Thus, gathering the most diverse data possible represents a core principle of this work, leading to two main strategies:

Category	Dataset
Video Classification	Kinetics710 [9, 101], Something-Something-v2 [25], ActivityNet [8], HACS Segment [130], COIN [90], HIREST [122], FineAction [63], PortraitMode-400 [31]
Temporal Action Localization	ActivityNet [8], HACS Segment [130], FineAction [63], Ego4D-MQ [26], COIN [90], HIREST [122], Perception-Test [79]
Video Temporal Grounding	Charade-STA [23], QVHighlight [48], Ego4D-NLQ [26], Didemo [32], QueryD [75], MedVidQA [29], Youcook2 [132], FineVideo [22], ActivityNet [8], HACS Segment [130], FineAction [63], Ego4D-MQ [26], COIN [90], HIREST [122], Perception-Test [79], EgoExoLearn [37]
Dense Video Captioning	ActivityNet [8], Youcook2 [132], EgoExoLearn [37], ViTT [35], HIREST [122], COIN [90]
Temporal Segmentation	Breakfast [46], ViTT [35]
Temporal Reasoning	ActivityNet-RTL [34]
General Video QA	TVQA [47], CLEVRER [115], NextQA [108], SportsQA [51], LLaVA-Video [128], FineVideo [22], VideoGPT+ [69], Oops [21], Perception-Test [79], EgoTaskQA [42], CinePile [82], STAR [104]
Multi-Page Document	SlideVQA [89], DUDE [96], MP-DocVQA [94]
Video Captioning	ActivityNet [8], Youcook2 [132], Shot2story [30], Vript [112], LLaVA-Video [128], Momentos [99], FunQA [109], S-MiT [74], LLaVA-Hound [127], Ego4D-HCap [41], EgoExoLearn [37]
Long Text	LongAlign [6], LongReward [125]

Table 1: Video, multi-page document, and long text dataset used in Eagle-2.5.

- *Human-annotated Data:* We integrate various open-source human-annotated datasets, including established video and image-document collections such as COIN [90] and SlideVQA [89], which can be directly considered as high-quality data.
- *Synthetic Video Data:* Considering that videos naturally contain long-context information, we incorporate open-source synthetic video data, such as LLaVA-Video [128]. These datasets are primarily annotated automatically using state-of-the-art models including GPT-4V/4o [76, 77], Claude-3 [2], and Gemini-1.5 Pro [83].

Combined with short-context data, all collected open-source datasets are summarized in Tab. 1. For convenience, we refer to this collective dataset as Open-Data.

3.3.2 Eagle-Video-110K

We curate Eagle-Video-110K to enhance long video understanding capabilities. Specifically, we first collect videos using a diversity-driven strategy. We then automatically annotate these videos using both *top-down* and *bottom-up* approaches to generate comprehensive story-level and fine-grained clip-level annotations, as shown in Fig. 5.

Diversity-driven video collection. We utilize several data sources for our video collection: Vidchapters [111], MiraData [44], InternVid-10M [100], Panda-70M [14], Vript [112], Shot2story [30], ViTT [35], and WebVid-10M [7], collectively referred to as A. Our approach prioritizes diversity, focusing on gathering a wide range of video content. For the current training dataset B, we use CLIP [81] to extract temporal features at a rate of 1 frame per second. Videos from both A and B are segmented into 10-second clips. We perform a pooling operation on each clip's frames to derive a representative feature vector. Let $\{b_i\}_{i=1}^{N_B}$ represent the clips from B, and $\{a_j\}_{j=1}^{N_A}$ represent those from A. We calculate the pairwise cosine similarity between clips from B and A. For each clip a_j in A, we identify its maximum similarity with any clip in B: $S_{\max}(a_j) = \max_{1 \le i \le N_B} S(b_i, a_j)$ We then introduce a similarity threshold $\tau = 0.5$. Clips in A with $S_{\max}(a_j)$ below this threshold are considered most novel relative to B: $A_{\text{novel}} = \{a_j \in A \mid S_{\max}(a_j) < \tau\}$

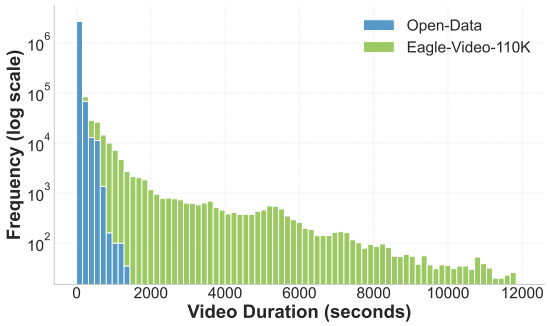

Figure 4: Comparison of video duration between open-source data and Eagle-Video-110K.

The clips in A_{novel} and their original videos are selected to enhance the diversity of our collection.

Story-level video data. We construct story-level annotations for long videos using a *top-down* approach. Unlike existing video datasets such as Shot2story [30], which employs shot detection to segment videos and construct storylines across shots, our methodology differs fundamentally. Shot-level segmentation often results in over-segmentation, producing excessively detailed annotations that are suboptimal for constructing coherent story-level text. Instead, we leverage human-annotated chapters as video segments, which provide more semantically meaningful annotations. We incorporate

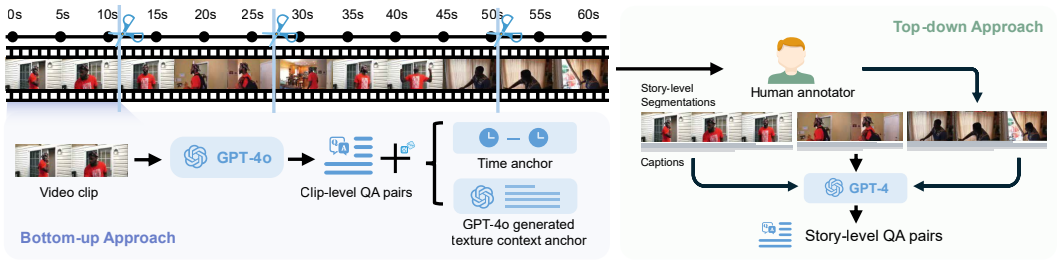

Figure 5: **Overview of our video annotation framework combining bottom-up clip-level and top-down story-level approaches.** The diagram illustrates our dual annotation strategy. In the bottom-up approach (left), short video clips are processed by GPT-4o to generate clip-level QA pairs enhanced with time anchors and textural context anchors. In the top-down approach (right), human annotators create story-level segmentations of longer videos, which are then captioned and processed by GPT-4 to generate comprehensive story-level QA pairs. This hierarchical methodology enables both fine-grained temporal understanding and high-level semantic comprehension of video content.

content from ViTT [35] and Vidchapters [111] among the selected videos and filter out any videos with fewer than two chapters to ensure they serve as effective story-level sources.

- *Chapter-level dense caption.* For a video divided into N clips, where each clip spans from timestamp a to b, we perform visual captioning for each segment individually. For each segment, frames are sampled at a rate of up to 2 frames per second, with a maximum of 50 frames. These sampled frames, together with user-provided segment titles, guide GPT-4o [77] in generating detailed visual descriptions focused on the content indicated by the titles.
- *Long-form QA generation.* Once visual descriptions for all segments are completed, we compile the captions for the entire video along with their corresponding time intervals and chapter titles. This aggregated information is provided to GPT-4 [1], which generates diverse question-answer pairs covering multiple question types.

Clip-level video data. Story-level video data typically emphasizes high-level semantic information that unfolds over extended periods. However, for general queries, it is often necessary to focus on localized spatiotemporal details. To address this need, we propose a *bottom-up*, computationally efficient automatic annotation method. This approach enables the generation of short clip annotations and facilitates the conversion of segment-level annotations into video-level ones by incorporating temporal and contextual anchors.

- *Clip-level video QA generation.* We generate QA pairs for each short clip in dataset A based on various question types. Specifically, we sample frames from each short clip at a rate of up to 2 frames per second and input them into GPT-4o. From a predefined question type pool, we randomly select five question types and prompt the model to generate corresponding question-answer pairs.
- *Clip-to-video QA conversion.* Since annotations for individual clips are designed for localized queries, conflicts may arise when these queries are extended to the entire video. To address this issue, we introduce two types of anchors for each clip-question pair: (1) We directly incorporate time intervals into questions to establish temporal references; (2) Using GPT-4o, we generate textual context anchors that provide additional information without revealing the answers.

4 Experiments

4.1 Comparison with State-of-the-Art VLMs

Video benchmarks. As shown in Tab. 2, Eagle2.5-8B demonstrates strong performance across multiple video understanding benchmarks. It achieves 74.8 on MVBench [54], 82.0 on Perception_test [78] and 72.2 on EgoSchema, outperforming similar-sized models like InternVL2.5-8B [92] (72.0, -, -) and Qwen2.5-VL-8B [97] (69.6, 70.5, 65.0). Eagle2.5-8B excels in MLVU [131] (77.6) and LongVideobench [106] (66.4), surpassing even InternVL2.5-78B (75.7, 63.6). For VideoMME (w/o subtitle), the performance of Eagle 2.5 (72.4) significantly surpasses models of the same size and is extremely close to the 72B parameter model. On CG-Bench [11], it scores 55.8, 46.6, 45.6, 13.4 across metrics, exceeding Claude-3.5-Sonnet [2] (56.5, 40.3, 35.6, 4.17) and Gemini-1.5-Pro [83]

Model	MVBench	Perception_test	EgoSchema	MMB-Video	MLVU	LVBench	Video-MME		CG-Bench				HourVideo		Charade-STA
	-	Val	fullset	-	Val	Val	w/o subtitle	w subtitle	Clue	Long	Open	mIoU	Dev	Test	mIoU
Closed-Source Models															
GPT-4o-0806 [77]	-	-	-	1.63	-	66.7	71.9	77.2	58.6	44.9	39.2	5.73	-	-	35.7
Claude-3.5-Sonnet [2]	-	-	-	-	-	-	60.0	62.9	56.5	40.3	35.6	4.17	-	-	-
Gemini-1.5-Pro [83]	-	-	72.2	1.30	-	64.0	75.0	81.3	50.9	37.8	28.7	3.85	37.2	37.4	-
Gemini-2.5-Pro [19]	-	-	-	-	81.2	69.2	87.0	-	-	-	-	-	-	-	-
Seed1.5-VL [28]	74.3	-	-	-	81.8	64.6	77.6	-	-	-	-	-	-	-	64.7
Publicly Available Models															
MiniCPM-V2.6-8B [113]	-	-	-	-	-	-	60.9	63.7	44.4	29.9	26.3	2.27	-	-	-
LongVILA-8B [15]	67.1	58.1	67.7	-	-	57.1	60.1	65.1	47.5	34.3	26.6	-	-	-	-
InternVL2.5-8B [17]	72.0	-	-	1.68	68.9	60.0	64.2	66.9	-	-	-	-	-	-	-
LLaVA-Video-8B [129]	58.6	67.9	57.3	-	70.8	58.2	63.3	69.7	-	-	-	-	-	-	-
Qwen2.5-VL-8B [5]	69.6	70.5	65.0	1.79	70.2	56.0	65.1	71.6	44.5	35.5	24.1	2.48	-	-	43.6
VideoChat-Flash-8B [56]	74.0	76.2	-	-	74.6	64.7	65.3	69.7	52.8	43.1	37.5	1.49	-	-	-
InternVL2.5-78B [17]	76.4	-	-	1.97	75.7	63.6	72.1	74.0	59.5	44.2	34.2	3.90	-	-	-
Qwen2.5-VL-72B [5]	70.4	73.2	76.2	2.02	74.6	60.7	73.3	79.1	-	-	-	-	-	-	50.9
LLaVA-Video-72B [129]	64.1	74.3	65.6	-	74.4	61.9	70.6	76.9	-	-	-	-	-	-	-
Eagle2.5-8B	**74.8**	**82.0**	**72.2**	**1.94**	**77.6**	**66.4**	**72.4**	**75.7**	**55.8**	**46.6**	**45.6**	**13.4**	**44.5**	**41.8**	**65.9**

Table 2: **Comparison with SoTA models on Various Video Benchmarks**. We sample each video at 2 FPS by default and disable tiling, and limit the minimum sampling frame number to 8 frames. Among them, the maximum frame number of Video-MME is 512, and the others are 256. Perception-Test turns on tiling to enable high-resolution testing.

Model	DocVQA	ChartQA	InfoVQA	TextVQA	OCRBench	MMstar	RWQA	AI2D	MMMU	$MMB_{1.1}$	MMVet	HallB	MathVista	**Avg**
	Test	Test	Test	Val	Test	Test	Test	Test	Val	Test	Test	Test	Test-Mini	**Score**
Closed-Source Models														
GPT-4o-0806 [77]	92.8	85.7	79.2	77.4	736	64.7	75.4	84.6	69.1	83.1	69.1	55.0	63.8	74.9
Claude-3.5-Sonnet [2]	95.2	90.8	74.3	74.1	788	65.1	60.1	81.2	68.3	80.9	70.1	55.5	67.7	74.0
Gemini-1.5-Pro [83]	93.1	87.2	81.0	78.8	754	59.1	67.5	79.1	62.2	74.6	64.0	45.6	63.9	71.7
Publicly Available Models														
MiniCPM-V2.6-8B [113]	90.8	82.4	-	80.1	852	57.5	65.0	82.1	49.8	78.0	60.0	48.1	60.6	-
LLaVA-One-Vision-8B [49]	87.5	80.0	68.8	-	622	61.7	66.3	81.4	48.8	80.9	57.5	31.6	63.2	-
InternVL2.5-8B [17]	93.0	84.8	77.6	79.1	822	62.8	70.1	84.5	56.0	**83.2**	62.8	50.1	64.4	73.1
Qwen2.5-VL-8B [5]	**95.7**	87.3	**82.6**	**84.9**	864	63.9	68.5	83.9	**58.6**	82.6	**67.1**	52.9	**68.2**	75.6
LLaVA-One-Vision-72B [49]	91.7	83.7	74.9	-	741	66.1	71.9	85.6	56.6	84.5	60.6	47.5	68.4	-
LLaMa-3.2-90B-Vision [20]	90.1	85.5	-	-	783	55.3	-	-	60.3	77.3	64.1	44.1	57.3	-
Eagle2.5-8B	94.1	**87.5**	80.4	83.7	**869**	**66.2**	**76.7**	**84.5**	55.8	81.7	62.9	**54.7**	67.8	**75.6**

Table 3: **Comparison with SoTA models on Various Image Benchmarks.** The avg score is computed as the average of all benchmark scores, with OCRBench score divided by 10.

(50.9, 37.8, 28.7, 3.85). With 44.5 on HourVideo [10] dev set and 41.8 on test set, all surpassing Gemini-1.5-Pro [83]. Finally, on Charade-STA [23], Eagle 2.5 outperforms other models significantly, demonstrating strong temporal perception capabilities. Eagle2.5-8B shows effective long-form video understanding, highlighting its robust visual reasoning using less parameters.

Image benchmarks. As shown in Tab. 3, Eagle2.5-8B demonstrates competitive performance across diverse image understanding benchmarks. It achieves strong results on document understanding (94.1 on DocVQA [71]), chart interpretation (87.5 on ChartQA [70]), and general information extraction (80.4 on InfoVQA [72], 83.7 on TextVQA [86]). The model also performs well in optical character recognition with 869 on OCRBench [65], comparable to other models in its category. Eagle2.5-8B shows balanced capabilities across multimodal general perception and reasoning tasks, scoring 66.2 on MMstar [12], 76.7 on RWQA [107], and 81.7 on $MMB_{1.1}$ [64], and 62.9 on MMVet [118]. Its performance extends to knowledge domain (55.8 on MMMU [119], 84.5 on AI2D [33]), visual hallucination benchmark (54.7 on HallB [27]), and mathematical reasoning (67.8 on MathVista [68]). Overall, Eagle2.5-8B achieves a competitive 75.6 average score, demonstrating its effectiveness as a versatile vision-language model that balances performance across various visual understanding tasks.

4.2 Ablation Studies

In this section, we conduct experiments on various benchmarks to evaluate our method. We mainly design experiments to study the following questions.

Q1: How do video and image data influence each other's benchmarks? Tab. 4 studies the impact of long context data on the image benchmark performance. We compare the image benchmark performance without training with long-context data and with training long-context data under different $\mathcal{L}_{\max}$. The results show that increasing the long-context data, under our training recipe, does not harm the short-context images and even slightly benefits it. To assess the impact of image data and pre-training on video benchmarks, we conduct a comparison using the $\mathcal{L}_{\max} = 32K$ model.

Training & Data recipe	DocVQA Val	ChartQA Test	InfoVQA Val	TextVQA Val	OCRBench Val	MMstar Test	RWQA Test	AI2D Test	MMMU Val	$MMB_{1.1}$ EN-Val	MMVet Test	HallB Test	MathVista Test-Mini	**Avg Score**
Eagle2.5-S2	92.6	88.3	78.8	84.6	868	66.5	74.4	85.5	54.0	85.5	57.3	53.4	65.1	74.8
Eagle2.5-S2+Eagle2.5-S2, $\mathcal{L}_{\max} = 32K$	92.3	86.6	77.6	82.8	861	66.7	75.9	83.7	55.5	84.8	63.6	55.4	68.3	75.3
Eagle2.5-S2+Eagle2.5-S2, $\mathcal{L}_{\max} = 64K$	92.5	87.0	78.4	83.9	865	66.8	76.8	83.9	55.7	85.2	63.3	55.2	67.3	75.6
Eagle2.5-S2+Eagle2.5-S2, $\mathcal{L}_{\max} = 128K$	93.2	87.5	78.5	83.7	869	66.2	76.7	84.5	55.8	85.5	62.9	54.7	67.8	75.7

Table 4: Impact of long-context data on performance of image benchmarks.

Training & Data recipe	MVBench -	MLVU Val	Video-MME w/o subtitle
S1→S2	70.4	67.4	64.9
S1→S1.5→S2 (OD+EV-110K)	72.9	70.9	65.2
S1→S1.5→S2 (Image+OD+EV-110K)	73.1	71.5	65.4

Table 5: The impact of image data and pretraining on the performance of video benchmarks. S1/S1.5 denotes the stage-1 and stage-1.5 similar to Eagle2 [3]. "OD" is short for Open-Data.

Recipe	InfoVQA Val	DocVQA Val	TextVQA Val	PT Val	MLVU Val	Video-MME w/o subtitle
baseline	77.6	92.3	82.8	76.3	71.5	65.4
w/o IAP	76.2	91.9	82.4	73.3	71.2	64.9
w/o ADS	77.0	92.1	82.8	75.5	70.1	65.0

Table 6: The impact of information-first sampling on performance of image and video benchmarks. The baseline is equipped with IAP and ADS strategy. "PT" is short for PereptionTest.

For each benchmark, we sampled at 2FPS, ensuring a maximum of 32 frames. As shown in Tab. 5, extensive image pre-training significantly enhances performance on short video benchmarks like MVBench, as well as on the relatively simple long video benchmark, MLVU. However, for the more challenging and held-out long video benchmark, Video-MME, the improvements are less pronounced.

Q2: The effect of information-first sampling on performance? Tab. 6 illustrates the impact of the information-first sampling strategy on image and video tasks. Without the Image Area Preservation strategy, high-resolution image benchmarks like InfoVQA and fine-grained video benchmarks such as Perception-test suffer significant performance degradation. The effect on other benchmarks is less pronounced. While the Automatic Degradation Sampling strategy offers convenience for processing various visual inputs, experiments indicate that omitting it poses a risk. The vision-context-centric strategy may truncate supervision signals, leading to performance loss.

Q3: The impact of different post-training schedules? Tab. 7 illustrates the performance impact of progressive mixed training from 32K to 64K compared to direct 64K mixed training on the video benchmarks. The results demonstrate that progressive training outperforms direct 64K mixed training, possibly due to two reasons: 1) Direct 64K hybrid training disperses samples across the 64K space, diluting the focus on shorter contexts. 2) Some longer samples are challenging to learn without a gradual learning process that transitions from easy to difficult. Fig. 6 shows the effect of progressive mixed training on the Video-MME benchmark. It reveals that as progressive training advances, the model's capacity to process more frames is gradually enhanced.

Training & Data recipe	MVBench -	MLVU Val	Video-MME w/o sbutitle
32K→64K, Open-Data	73.0	74.5	68.1
64K, Open-Data	71.3	74.0	67.9
32K→64K, Open-Data+EV-110K	73.9	75.1	68.8

Table 7: The impact of Eagle-Video-110K dataset and different post-training schedules on the performance of video benchmarks.

Q4: The impact of Eagle-Video-110K data on performance? We assess the impact of Eagle-Video-110K on model performance. Table 7 shows that it improves several long and short video benchmarks. Figure 6 demonstrates that training with Eagle-Video-110K enhances the model's ability to process many frames ($\geq$ 128 frames) by incorporating long videos absent from the open-source training dataset.

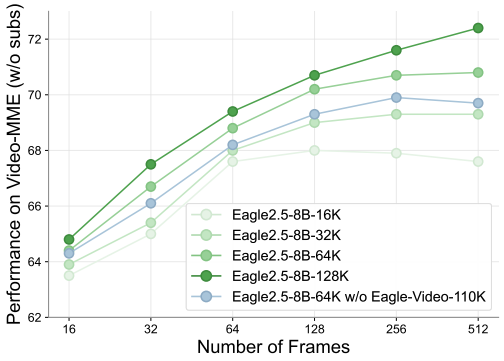

Figure 6: The impact of Eagle-Video-110K dataset and different post-training schedules on the performance of Video-MME.

5 Conclusion

In this work, we present Eagle 2.5, an advanced vision-language model family designed for long-context multimodal understanding. Through innovative training approaches - including information-first sampling and progressive mixed post-training - combined with our dual-annotated Eagle-Video-110K dataset, we significantly enhance long-context

comprehension capabilities. Eagle 2.5 achieves leading performance on video and high-resolution image benchmarks, matching larger models like GPT-4V and Gemini 1.5 Pro despite its smaller size. With advanced training strategies and diverse data, Eagle 2.5 sets a strong foundation for future research, paving the way for efficient and versatile VLMs in complex real-world scenarios.

Limitations. The training of Eagle2.5 required substantial computational resources, specifically a cluster of 128 H100 GPUs. This high resource demand may limit the reproducibility and accessibility of our approach for researchers or practitioners without access to large-scale GPU infrastructure. Future work could explore more resource-efficient training strategies or model architectures to reduce computational requirements.

6 Acknowledgement

This work is supported by the National Key R&D Program of China (No. 2022ZD0160900), Jiangsu Frontier Technology Research and Development Program (No. BF2024076), the National Natural Science Foundation of China (Grant No.62372223 and U24A20330), and in part by Nanjing University-China Mobile Communications Group Co., Ltd. Joint Institute under Grant NJ20250037.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv:2303.08774*, 2023.

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www.anthropic.com, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[3] Anonymous Authors. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *In Submission*, 2025.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv:2308.12966*, 2023.

[5] 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:2502.13923*, 2025.

[6] Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. *arXiv:2401.18058*, 2024.

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[10] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. In *NeurIPS*, 2025.

[11] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv:2412.12075*, 2024.

[12] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv:2403.20330*, 2024.

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

[14] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024.

[15] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv:2408.10188*, 2024.

[16] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv:2312.14238*, 2023.

[17] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv:2412.05271*, 2024.

[18] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NVLM: Open frontier-class multimodal llms. *arXiv:2409.11402*, 2024.

[19] Google DeepMind. Build rich, interactive web apps with an updated gemini 2.5 pro. https://blog.google/products/gemini/gemini-2-5-pro-updates/, 2025. Accessed: 2025.

[20] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024.

[21] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *CVPR*, 2020.

[22] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

[23] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017.

[24] Ridouane Ghermi, Xi Wang, Vicky Kalogeiton, and Ivan Laptev. Long story short: Story-level video understanding from 20k short films, 2025. URL https://arxiv.org/abs/2406.10221.

[25] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[27] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv:2310.14566*, 2023.

[28] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[29] Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158, 2023.

[30] Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv:2312.10300*, 2023.

[31] Mingfei Han, Linjie Yang, Xiaojie Jin, Jiashi Feng, Xiaojun Chang, and Heng Wang. Video recognition in portrait mode. In *CVPR*, 2024.

[32] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv:2003.10286*, 2020.

[33] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688, 2021.

[34] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. LITA: Language instructed temporal-localization assistant. In *ECCV*, 2024.

[35] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. *CoRR*, abs/2011.11760, 2020.

[36] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, 2020.

[37] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *CVPR*, 2024.

[38] Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Lijin Yang, Xinyuan Chen, Yaohui Wang, Zheng Nie, Jinyao Liu, et al. Vinci: A real-time embodied smart assistant based on egocentric vision-language model. *arXiv preprint arXiv:2412.21080*, 2024.

[39] Yifei Huang, Jilan Xu, Baoqi Pei, Lijin Yang, Mingfang Zhang, Yuping He, Guo Chen, Xinyuan Chen, Yaohui Wang, Zheng Nie, et al. Vinci: A real-time smart assistant based on egocentric vision-language model for portable devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–33, 2025.

[40] IDEFICS. Introducing IDEFICS: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics, 2023.

[41] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *CVPR*, 2024.

[42] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In *NeurIPS*, 2022.

[43] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, 2024.

[44] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In *NeurIPS*, 2025.

[45] Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *ECCV*, 2024.

[46] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. 2014.

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

[48] Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. 2021.

[49] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv:2408.03326*, 2024.

[50] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024.

[51] Haopeng Li, Andong Deng, Qiuhong Ke, Jun Liu, Hossein Rahmani, Yulan Guo, Bernt Schiele, and Chen Chen. Sports-qa: A large-scale video question answering benchmark for complex and professional sports. *arXiv:2401.01505*, 2024.

[52] 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 *ICML*, 2023.

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

[54] 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 *CVPR*, 2024.

[55] Pengyi Li, Irina Abdullaeva, Alexander Gambashidze, Andrey Kuznetsov, and Ivan Oseledets. Maxinfo: A training-free key-frame selection method using maximum volume for enhanced video understanding. *arXiv:2502.03183*, 2025.

[56] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

[57] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*. Springer, 2024.

[58] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[59] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

[60] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[61] Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. *arXiv preprint arXiv:2505.14454*, 2025.

[62] Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Bo Zheng, Linfeng Zhang, Siteng Huang, and Honggang Chen. Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. *arXiv preprint arXiv:2501.05179*, 2025.

[63] Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31:6937–6950, 2022.

[64] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

[65] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[66] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. NVILA: Efficient frontier visual language models. In *CVPR*, 2025.

[67] Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms, 2025. URL https://arxiv.org/abs/2506.05328.

[68] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*, 2023.

[69] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv:2406.09418*, 2024.

[70] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022.

[71] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A dataset for vqa on document images. In *WACV*, 2021.

[72] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *WACV*, 2022.

[73] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[74] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *CVPR*, 2021.

[75] Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP*, 2021.

[76] OpenAI. Gpt-4v(ision) system card, 2023. URL https://openai.com/index/gpt-4v-system-card/.

[77] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2023. Accessed: 2024-11-12.

[78] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2023.

[79] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2023.

[80] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. In *NeurIPS*, 2025.

[81] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PmLR, 2021.

[82] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv:2405.08813*, 2024.

[83] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

[84] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv:2410.17434*, 2024.

[85] Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Xiawu Zheng, Yan Zhang, et al. Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuray. *arXiv:2502.05177*, 2025.

[86] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

[87] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024.

[88] Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. Moviellm: Enhancing long video understanding with ai-generated movies. *arXiv:2403.01422*, 2024.

[89] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, 2023.

[90] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, 2019.

[91] 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:2312.11805*, 2023.

[92] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. https://internvl.github.io/blog/2024-07-02-InternVL-2.0/, 2024.

[93] Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

[94] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

[95] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv:2406.16860*, 2024.

[96] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *ICCV*, 2023.

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

[98] Shihao Wang, Guo Chen, De-An Huang, Zhiqi Li, Minghan Li, Guilin Liu, Jose M. Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal grounding. *arXiv preprint arXiv:2507.13353*, 2025.

[99] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv:2401.10529*, 2024.

[100] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024.

[101] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. InternVideo2: Scaling video foundation models for multimodal video understanding. *arXiv:2403.15377*, 2024.

[102] Yuxuan Wang, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long-context video understanding with recurrent memory bridges. *arXiv:2409.01071*, 2024.

[103] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *ECCV*. Springer, 2024.

[104] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. *CoRR*, abs/2405.09711, 2024.

[105] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *CVPR*, 2021.

[106] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*, 2025.

[107] X.ai. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, 2024.

[108] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. 2021.

[109] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *ECCV*, 2024.

[110] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv:2408.10188*, 2024.

[111] Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vidchapters-7m: Video chapters at scale. In *NeurIPS*, 2023.

[112] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In *NeurIPS*, 2025.

[113] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv:2408.01800*, 2024.

[114] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8579–8591, 2025.

[115] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *ICLR*, 2020.

[116] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023.

[117] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *arXiv:2410.03226*, 2024.

[118] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv:2308.02490*, 2023.

[119] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

[120] Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. Movie101: A new movie understanding benchmark. *arXiv:2305.12140*, 2023.

[121] Zihao Yue, Yepeng Zhang, Ziheng Wang, and Qin Jin. Movie101v2: Improved movie narration benchmark. *arXiv:2404.13370*, 2024.

[122] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *CVPR*, 2023.

[123] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[124] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv:2406.08085*, 2024.

[125] Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. Longreward: Improving long-context large language models with ai feedback. *arXiv:2410.21252*, 2024.

[126] 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:2406.16852*, 2024.

[127] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv:2404.01258*, 2024.

[128] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024.

[129] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024.

[130] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV*, 2019.

[131] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv:2406.04264*, 2024.

[132] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

[133] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main claims: Eagle2.5 is a frontier vision-language model (VLM) for long-context multimodal learning, introducing a generalist framework for both long video comprehension and high-resolution image understanding. The paper claims the introduction of two new techniques (Automatic Degrade Sampling and Image Area Preservation), a new dataset (Eagle-Video-110K), and substantial improvements on long-context benchmarks, with performance matching or surpassing top-tier commercial and open-source models. These claims are supported by the theoretical and experimental sections, and the scope is well-aligned with the results presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper does not present new theoretical results, formal theorems, or proofs. The work is empirical and methodological, focusing on architectural and data innovations, so this criterion does not apply.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the model architecture, training strategies, data recipes, and benchmark evaluation protocols. It specifies the datasets used, the training schedule (including context lengths), and the evaluation settings (e.g., sampling rates, frame counts). While code and data are not directly linked in the main text, the methods and procedures are described in sufficient detail for an expert to attempt reproduction of the main results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: There is no explicit statement in the main paper about releasing code or data, nor are URLs or instructions provided for accessing the Eagle2.5 model, Eagle-Video-110K dataset, or scripts to reproduce experiments. The paper summarizes data sources and contributions but does not provide open access or reproducibility instructions as supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides extensive details about the experimental setup, including data splits, sampling rates, frame counts, tiling parameters, and training schedules (e.g., progressive training with specified context lengths). Tables specify which datasets are used for which tasks, and ablation studies clarify the impact of different settings. Hyperparameters and optimizer choices are not exhaustively listed, but the level of detail is sufficient for understanding and contextualizing the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports benchmark scores and ablation results in tables but does not include error bars, confidence intervals, or statistical significance tests for the reported results. There is no discussion of variability, standard deviation, or the statistical robustness of the findings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper does not provide information about the compute resources used for training or evaluation (e.g., type of GPUs, number of nodes, training time, or memory requirements). This information is absent from both the main text and tables.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: There is no indication that the research violates the NeurIPS Code of Ethics. The paper uses publicly available datasets or those curated by the authors, and there is no evidence of unethical practices in data collection, model training, or evaluation.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper does not include a discussion of broader societal impacts, either positive or negative. There is no section or paragraph addressing potential misuse, fairness, privacy, or security implications of releasing a powerful long-context VLM or dataset.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper does not describe any safeguards for responsible release of the model or dataset, nor does it discuss filtering for unsafe content or mechanisms to prevent misuse of the released assets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces Eagle-Video-110K as a new dataset but does not provide documentation, access instructions, or details about licensing, consent, or terms of use for the new asset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.