VisCount: Evaluating Counting Capabilities of Multimodal Video-LLMs

Swetha Krishnan
University of Massachusetts
Amherst, MA
swethakrishn@umass.edu

Rahasya Barkur
University of Massachusetts
Amherst, MA
rbarkur@umass.edu

Kushal Raju
University of Massachusetts
Amherst, MA
kraju@umass.edu

Abstract

The rise of multimodal Video-LLMs, driven by advancements in Large Language Models (LLMs), has enabled powerful capabilities in video understanding tasks such as spatio-temporal reasoning, narrative comprehension, and video-based Visual Question Answering (VQA). However, despite the development of evaluation benchmarks for Video-LLMs, most existing efforts focus on complex reasoning tasks, overlooking evaluation on fundamental vision tasks like object counting. In this paper, we present VisCount, a novel dataset that comprehensively assesses the performance of multimodal Video-LLMs across low object frequencies. We evaluate 5 recent models, including both open-source and closed-source variants, and find that most of the Video-LLMs, especially open-source ones, struggle with counting accuracies over images and videos. Additionally, we find that performance of these models degrades with increasing object frequencies. Our findings provide valuable insights for robustness of video LLMs, revealing key performance gaps in video-based multimodal models.

1. Introduction

Visual understanding in computer vision has advanced significantly in recent years, with tasks becoming increasingly complex and modalities diversifying to include images, short videos, and long videos. Recent advances in video understanding have been largely driven by multimodal Video Large Language Models (Video LLMs), which aim to leverage the impressive contextual capabilities of LLMs with the help of vision encoders to improve Visual Question Answering (VQA) [6, 16]. Efforts in the research community have also led to the development of diverse video datasets, ranging from benchmarks for short video understanding like ActivityNet-QA [30] to datasets like EgoSchema [15], CinePile [19], and MovieChat-1K [25], which focus on the unique challenges of long video understanding, including character tracking, scene segmentation, and cross-episode reasoning.

Video understanding often necessitates the amalgamation of low-level, fundamental vision tasks such as activity recognition, object detection, and object counting. Among these, counting objects in videos stands out as a crucial elemental component, provide applications ranging from crowd analysis and traffic monitoring to object inventory and video understanding. It not only facilitates object detection but also provides contextual cues for temporal reasoning, as it captures frame-to-frame changes and enables models to reason about dynamic events and occlusions. For instance, in video datasets like MovieChat-1K [25] and CinePile [19], where models are expected to infer emotions and track character interactions over time, the ability to count objects (like people, vehicles, or objects of interest) provides essential contextual information. This contextual grounding enables more effective tracking, action prediction, and story comprehension. Previous research on object counting for images has primarily focused on two key paradigms: generic object counting and dense object counting. Generic object counting addresses the problem of counting objects of various categories in natural scenes, where the number of objects is typically small (fewer than 10) [1, 9]. On the other hand, dense object counting tackles scenarios with high-density object distributions, often seen in crowd counting datasets [24]. There thus exists a need for a comprehensive dataset for generic object counting which spans both images and videos, includes low-density and widely recognised objects, and captures temporal changes in video sequences.

To address these challenges, we introduce *VisCount*, a comprehensive dataset for more robust evaluations of multimodal video-LLMs, containing images and videos with varying object frequencies and object classes. Unlike existing datasets that emphasize dense object distributions or repetitive actions, VisCount aims to bridge the gap in evaluating the object counting capabilities of MultiModal-LLMs. Our main contributions are the following:

1. A new generic object counting dataset for low-density, well-separated objects and captures temporal changes in video sequences. *VisCount*, designed to assess count-

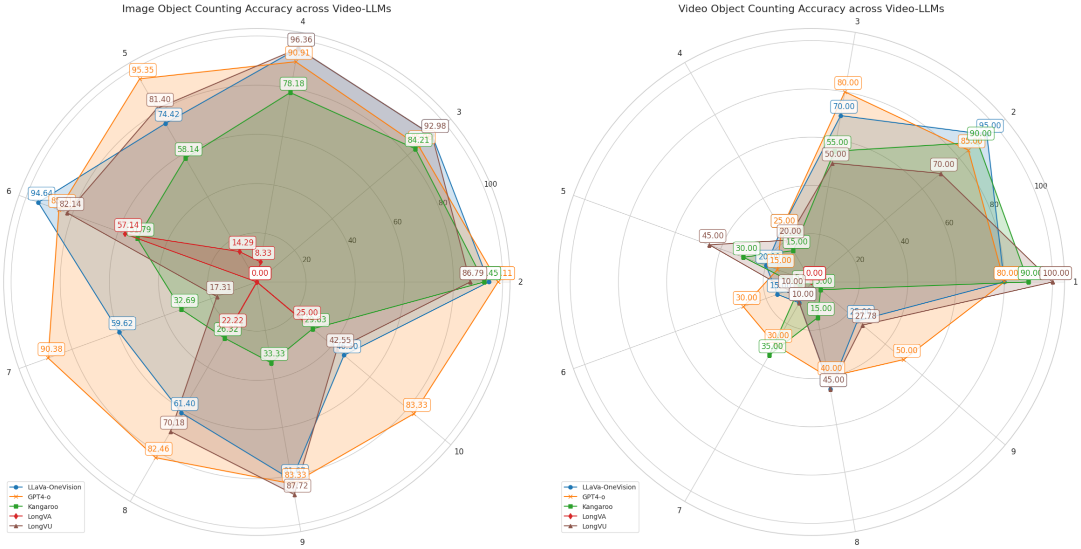


Figure 1. **Can Video-LMMs count objects accurately?** Classification accuracies on the proposed evaluation dataset, VisCount, for SoTA Video-LLMs over various object frequencies (1-10) across images and short videos. GPT-4o shows the best performance on images, while open-source models lag significantly in accuracy with higher object frequencies.

ing robustness of Multimodal LLMs across images and videos.

2. We evaluate both open-source and closed-source Video-LLMs on the dataset and find that most models show poor performance, revealing critical limitations in object counting, particularly at higher object frequencies.

2. Related Work

Video Understanding with LLM-based methods. Video-LLMs and large video multimodal models have emerged as a pivotal focus in the field of computer vision, driven by the growing demand for models capable of handling spatio-temporal reasoning [25], emotional comprehension [19], and complex narrative understanding [15]. With the growing availability and advancement of LLMs, the combination of pretrained vision encoders [18] with LLMs has become a dominant approach for developing powerful vision-language models (VLMs). The LLaVA series [7] and other works [20, 28] continue to push the boundaries through carefully curated datasets and advanced training techniques. LLaVA OneVision [6] pushes the fusion of vision and language even further, showcasing the potential of a unified model that comprehensively processes multimodal inputs. Kangaroo [10], an emerging framework, introduces novel methods for multimodal reasoning, offering improved efficiency and flexibility in handling complex vision-language tasks. LongVU [23] takes a unique approach by employing adaptive compression, allowing it to efficiently process real-world long videos. In contrast, LongVA [31] leverages its ability to handle unlimited context length, though this capability comes at a high computational cost, making it more resource-intensive for large-scale tasks.

Object Counting Datasets for images and videos. Object counting, which involves estimating the number of object instances in images or videos, is a long-standing problem in computer vision [1, 11]. Across images and videos, datasets of varying object classes [11, 26], and object densities ([24]) have been created for evaluating performance of object instance frequency estimation. Object counting in images involves generic object counting of everyday objects [1], or even indiscernible objects [26]. For videos, object counting is frequently framed as repetitive activity counting [32], where ”stuff” (e.g., actions) is counted instead of specific object instances. Additionally, datasets involving high object densities [24] are more commonly available than those for low frequencies involving good visibility of objects. Another critical gap emerges from the context of video-based Visual Question Answering (VQA). Questions such as, *”How many* $\{\texttt{objects}\}$ *do you see?”* are a common query type in VQA datasets [2], but explicit datasets dedicated to evaluating models on this form of counting are scarce.

Object Counting inaccuracies with LLMs. Despite their growing reputation for strong reasoning and inference capabilities, Large Language Models (LLMs) exhibit notable limitations in tasks requiring precise counting [29]. LLMs often miscount items in symbolic reasoning tasks or fail to correctly predict the number of repeated tokens in a text sequence [12, 13]. As a result, LLMs may conflate patterns with logical reasoning, leading to incorrect inferences about counts, repetitions, or sequences [14].

Dataset	Task	Object Type	Images	Video	#Images/Videos	Min./Max Freq.	Min/Max length (s)
Countix [3]	Activity Counting	Activities	✗	✓	- /8757	2/73	0.2/10
voc-18-bd-11 [11]	DOC	bird, blood cells	✗	✓	- /20	4/8	155/155
CountBench [17]	GOC (Descr.)	everyday obj.	✓	✗	540/ -	2/10	- / -
VisCount (Ours)	GOC	everyday obj.	✓	✓	491/181	1/10	1/161

Table 1. **Comparison with other datasets.** Statistics of existing datasets for object counting in images and videos. ’DOC’ stands for Dense Object Counting, and ’GOC’ stands for Generic Object Counting.

With the increasing adoption of LLM-based methods for vision tasks, this counting inaccuracy becomes even more concerning: unlike text-based counting, where tokenization is discrete, vision-based counting using LLM-based methods requires models to track objects across frames, handle occlusions, and differentiate between overlapping or visually similar objects. If counting errors persist from text to vision tasks, it could limit the reliability of LLMs for downstream applications like scene understanding, video VQA, and multi-object tracking. Thus, there is a need for targeted evaluation of object counting in LLM-based vision models.

Evaluation of Video-LLMs. Significant efforts have been made by the vision community to develop robust benchmarks for evaluating the performance of video-LLMs for long video (>30 minutes) and short video (<2 minutes) understanding [4, 5, 33]. MVBench [8] provides an early attempt at video-LLM evaluation, covering 20 video evaluation tasks for spatial and temporal understanding. Video-MME [4] introduces a multi-modal evaluation benchmark tailored for video analysis, comprising of 900 videos and 2,700 question-answer (QA) pairs for evaluating spatio-temporal reasoning, event understanding, and frame-based context extraction. For a more comprehensive evaluation, CVRR-ES [5] offers a broader testing framework to measure the reasoning, robustness, and interpretability of video-LLMs, with 217 videos and 2,400 open-ended QA pairs, for 11 distinct evaluation tasks, including multi-object tracking, motion prediction, and event recognition. Unlike simpler benchmarks, CVRR-ES emphasizes open-ended QA formats, rather than relying on predefined multiple-choice options. To address the challenges posed by long video understanding, MLVU [33] provides one of the most comprehensive benchmarks in the field, with over 1,300 videos and 2,500 evaluation tasks, for long-form video comprehension. MLVU incorporates segmentation, narrative tracking, and cross-episode reasoning, to assess for learning long-form reasoning.

3. Method

The goal is to create a dataset that focuses mainly on object frequency evaluation, where the objects are easily seen in the respective images and videos.

Challenges in object counting datasets. Overcoming these challenges demands the need for using automation tools for annotation, algorithms for object detection and segmentation, and scalable dataset collection pipelines.

3.1. Generating Image-QA pairs for object counting

A naive approach would be to filter a large-scale image datasets by considering only those examples in which the object frequency matches the segmentation labels for that class. Such an approach results in a noisy dataset, since objects in the image highly vary in occlusion, crowdedness, partial appearance, and scale. It is thus critical to ensure that the images contain generic objects with easily identifiable object classes, detectable in size and visibility, and should reflect object frequencies that are low enough to avoid dense appearances.

First, we took 540 images from CountBench [17] dataset, which was filtered down by Pass *et al.* (2023) from LAION-400M [21] dataset, containing 400 million CLIP-filtered image-text pairs. After filtration of non-existent images, we came down to 491 images, with up to 60 images per object frequency classes, ranging from 2 to 10. The CountBench annotations of these images provided general descriptions that were used to identify the object class and ground truth frequency. In some cases, the annotation did not provide a direct reference to the number of objects, thus requiring manual revisions and verification to generate the required annotations for the task. The questions were finally generated as ”How many {objects} are present?”, with the answers as ”{frequency} {objects}.”

3.2. Generating Video-QA pairs for object counting

Challenges in counting within videos are compounded by factors such as scale invariance, where objects appear at varying sizes due to their distance from the camera, and scene transitions, which cause abrupt changes that disrupt the flow of counting. Additionally, occlusions where objects are partially or fully blocked made it harder to track them accurately. Other challenges include reappearance and vanishing objects, where objects may briefly leave the frame and then reappear, or vanish altogether, complicating efforts to maintain an accurate count across dynamic scenes. We are considering these challenges while creating the datasets.

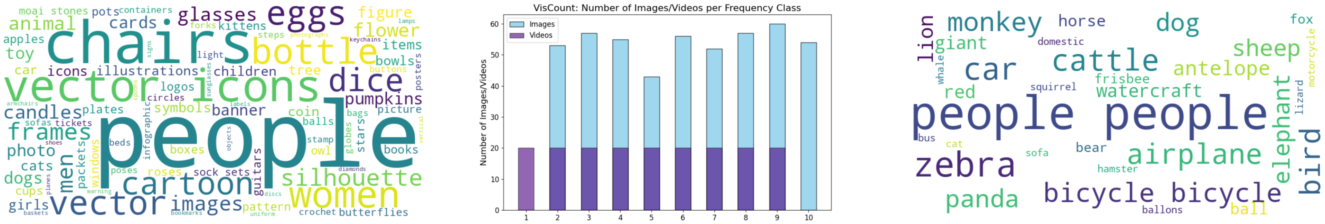


Figure 2. **VisCount Dataset Statistics.** Left:illustration of the most frequent object categories in the image dataset. Middle: Frequency distribution of objects. Right: illustration of the most frequent object categories in the video dataset.

We considered two prominent publicly available object detection datasets from video: ImageNet Video [22] and DanceTrack [27]. Non-people objects (e.g. animals) were extracted from the wide variety of objects in Image Net Video dataset. However, distribution of object counts the resultant video collection was highly uneven: the majority of videos contained only one object, with very few having more than five objects. This uneven distribution made it insufficient for our object-counting needs. To cover the remaining gaps, particularly in the higher object count ranges, we turned to the DanceTrack [27] dataset, which focuses on videos of various people dancing. DanceTrack provided more videos with higher object counts (people), allowing us to supplement the missing frequencies. By combining these two datasets, we were able to partially achieve the desired frequency distribution, though it still fell short of our goal.

Finally, to meet our target of at least 20 questions per object count frequency (up to 10 objects), we supplemented the dataset by extracting additional clips from YouTube. All of these clips were manually curated through Web scraping and annotated with the respective object-of-interest. The question template utilised was ”How many {objects} can be seen in the video?”, and the answer template followed was ”{frequency} {objects}.” This process aided in arriving at a balanced dataset, with a total of 175 videos addressing all object counts and ensuring better distribution across the frequencies, as depicted in Fig. 2.

3.3. Implementation Details

Models. We selectively curate a list of SOTA multimodal video-LLMs based on model size, visual encoding techniques and LLM backbone for evaluation on the generated dataset. We aim to use models that do not require prior sampling of frames to feed as video input to the model, i.e., the selected video-LLM should be able to process at least 250 frames in one pass. We thus arrive at five models, namely LLaVa-OneVision [6], Kangaroo [10], GPT-4 omni (GPT-4o) [16], LongVU[23], LongVA [31]. Tab. 2 provides an overview of the model parameters.

Image Evaluation. In the case of Kangaroo, which requires atleast 2 frames for evaluation, we feed each image as a 1-second video containing 3 frames, generated at 3fps.

Model	Size	LLM	Visual Encoder	#Frames	#Tokens
LLaVA-OneVision [6]	7B	Qwen2-7B	SigLIP-SO400M	1fps	7290
Kangaroo [10]	8B	LLaMA3-8B	EVA-CLIP-ViT-G/14	-	-
LongVU [23]	7.67B	Qwen2-7B	SigLIP + DINOv2	1fps	-
LongVA [31]	7B	Qwen2-7B-224K	CLIP-ViT-L/14	384	55,296
GPT-4o [16]	-	-	-	-	-

Table 2. Comparison of SoTA Video-LLMs chosen for evaluation.

Video Evaluation. GPT-4o’s API enforces a strict token-per-minute (TPM) limit of 30,000, which constrained our evaluation on the VisCount dataset. To comply, videos were sampled at every 50 frames before being processed. Despite this, six videos exceeded the token limit even after condensation, preventing their evaluation. For these cases, a predicted value of 0 was assigned.

4. VisCount: Vision-based Object Counting Evaluation Dataset

We introduce VisCount, a comprehensive dataset for generic object counting with images and short videos, curated from publicly available datasets and YouTube clips.

Overview. The image dataset contains a total of 491 images containing between two and ten object frequencies, annotated with respective question and ground truth answer. The video dataset consists of 180 questions spanning over 175 videos with 1 to 9 object frequencies. Average video length of 20.11 seconds, varying from as small as 1 second to 161.96 seconds. Fig. 2 provides additional visualisations of the image and video quantities and object categories available in this dataset. **Bias.** Additionally, it is important to note that as more images were filtered for this dataset, the result leads to a collection of imbalanced data, i.e. the number of images with lower frequencies (<5) are much higher than those with higher frequencies. As mentioned by Pass *et al.* (2023), it is worth noting that the dataset contains significant amount of images that are simple 2D objects (e.g. vector icons) in comparison with objects of everyday scenes. This is typical with general web-filtered images as well, given the difficulty of finding object frequencies greater than 10 that are not difficult to detect.

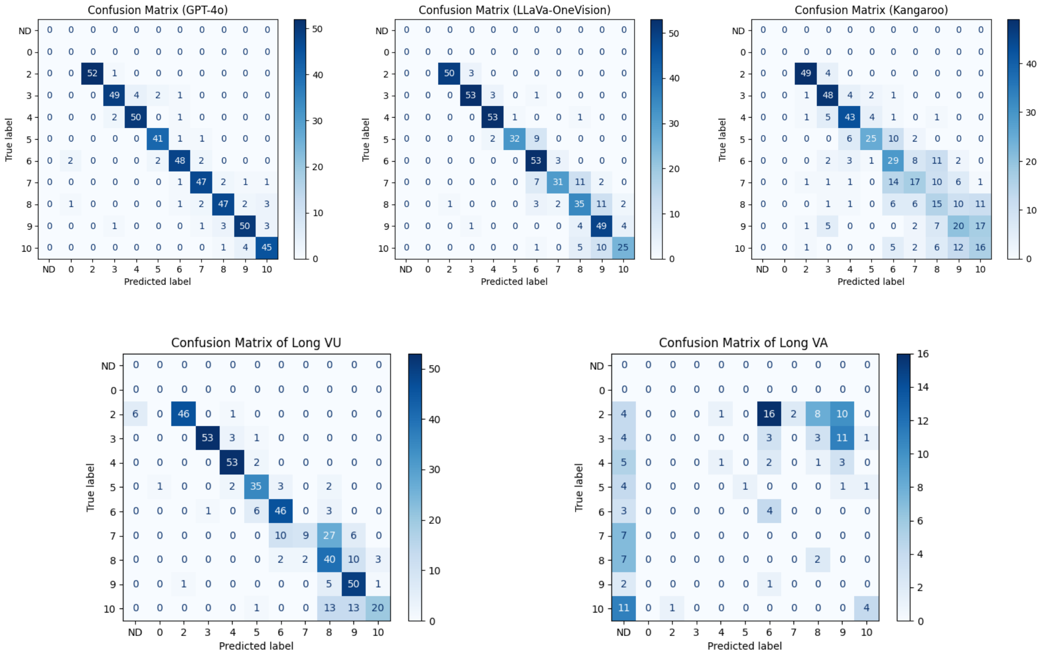


Figure 3. **Confusion matrices on VisCount images.** Classification accuracies on the proposed dataset, VisCount, broken down into confusion matrices over object frequencies for GPT-4o [16], LLaVa-OneVision [6], Kangaroo [10], LongVU [23] and LongVA [31] respectively. Results indicate clear incapabilites of Video-LLMs on generic object counting with images. 'ND' refers to cases where the model is unable to predict an exact count, and instead using words like 'several', 'multiple', etc.

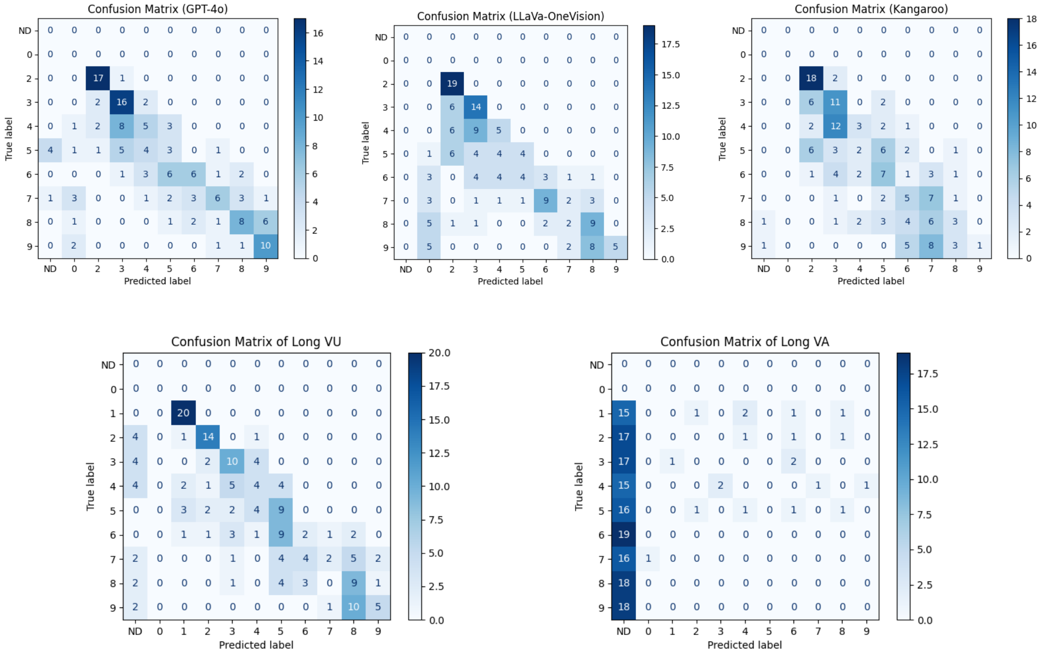
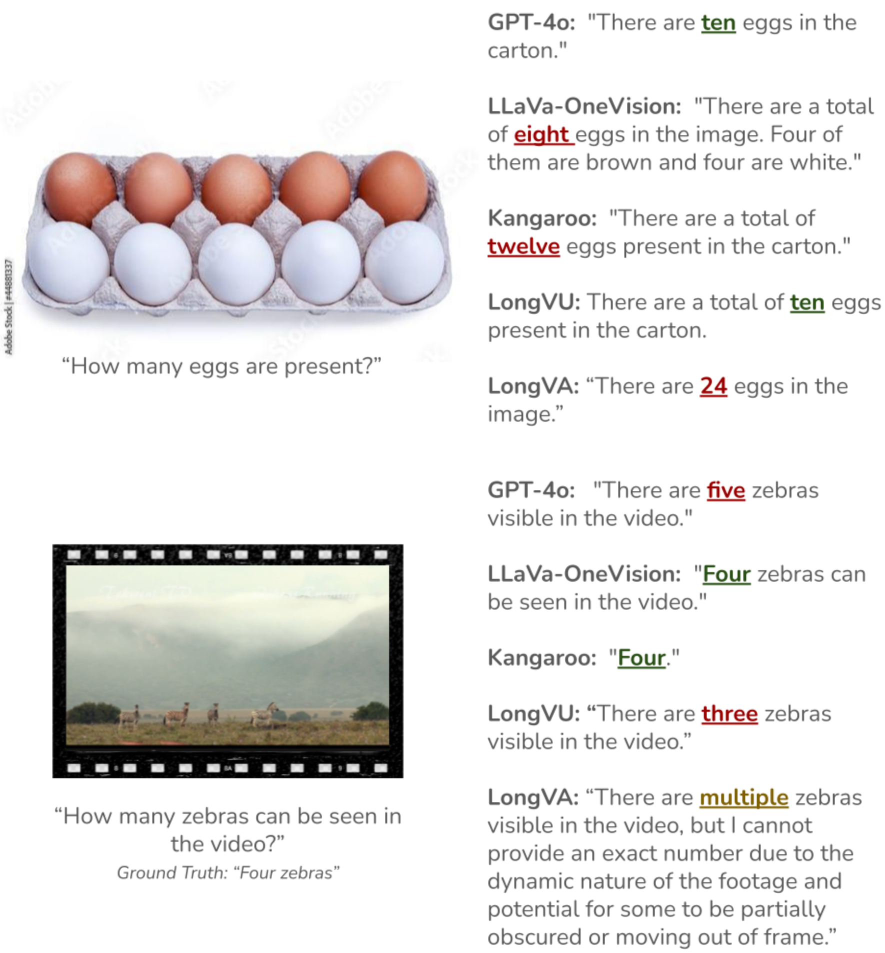


Figure 4. **Confusion matrices on VisCount videos.** Classification accuracies on the proposed dataset, VisCount, broken down into confusion matrices over object frequencies for GPT-4o [16], LLaVa-OneVision [6], Kangaroo [10], LongVU [23] and LongVA [31] respectively. Results indicate clear incapabilites of Video-LLMs on generic object counting with images. 'ND' refers to cases where the model is unable to predict an exact count, i.e. in cases where it estimates with words like 'several', 'multiple', etc.



"How many eggs are present?"

GPT-4o: "There are **ten** eggs in the carton."

LLaVa-OneVision: "There are a total of **eight** eggs in the image. Four of them are brown and four are white."

Kangaroo: "There are a total of **twelve** eggs present in the carton."

LongVU: There are a total of **ten** eggs present in the carton.

LongVA: "There are **24** eggs in the image."

"How many zebras can be seen in the video?"

Ground Truth: "Four zebras"

GPT-4o: "There are **five** zebras visible in the video."

LLaVa-OneVision: "**Four** zebras can be seen in the video."

Kangaroo: "**Four**."

LongVU: "There are **three** zebras visible in the video."

LongVA: "There are **multiple** zebras visible in the video, but I cannot provide an exact number due to the dynamic nature of the footage and potential for some to be partially obscured or moving out of frame."

Figure 5. **Example Cases.** Predicted answers from five multimodal Video-LLMs on a sample image and video from the VisCount dataset.

5. Experiments

5.1. Counting accuracy with images

Fig. 3 provides evaluation results of VisCount on GPT-4o [16], LLaVa-OneVision [6], Kangaroo [10], LongVU [23] and LongVA [31]. Example cases extracted with VisCount images and videos are highlighted in Fig. 5.

GPT-4o exhibits decent performance for lower counts (e.g., 0–3) but struggles as the object count increases, with the lowest accuracies (83%) observed for nine and ten objects. LLaVA-OneVision shows a similar trend to GPT-4o, but has more pronounced errors distributed across higher object counts. For LongVA, the ”ND” (Not Defined) category is prominently populated, which suggests that model often resorts to ambiguous or non-numeric descriptors like ”several” or ”multiple” rather than providing a specific count. Thus, these confusion matrices reveal that open-source multimodal video-LLMs are not well-suited for object counting and require improvement in numerical reasoning and object differentiation.

The comparatively weaker performance of open-source Video-LLMs may stem from multiple factors. First, the model size of commercial systems like GPT-4o likely dwarfs those of open-source counterparts, leveraging sheer scale to achieve superior representation and generalization. Second, commercial models are trained on vast and diverse datasets encompassing complex inter-dependencies across modalities, such as text, images, and audio, providing them with a broader foundation for numerical reasoning. Additionally, advanced architectural optimizations, including transformer-based enhancements and These factors collectively grant commercial models a significant edge in even simple tasks like object counting, which remains a critical challenge for open-source alternatives.

5.2. Counting accuracy with videos

Fig Fig. 1 presents the evaluation results of several models on the video subset of VisCount, including GPT-4o [16],

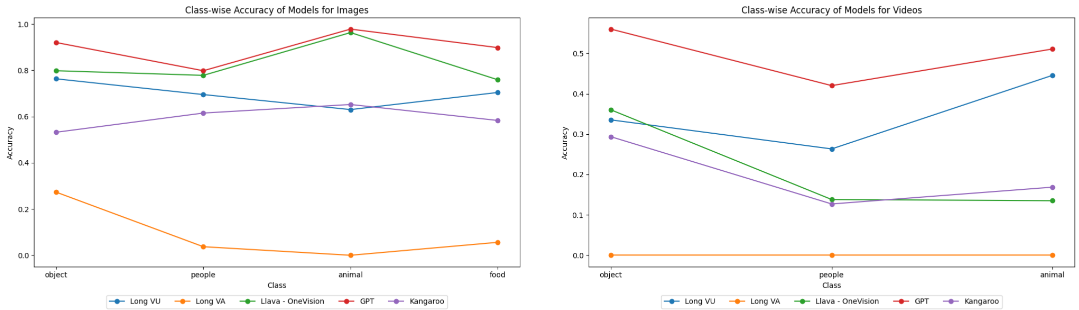


Figure 6. **Object-wise classification accuracies on the proposed dataset, VisCount**, segregated into 4 groups, namely *people, animal, food and object*, for GPT-4o [16], LLaVa-OneVision [6], Kangaroo [10], LongVU [23] and LongVA [31] respectively.

LLaVa-OneVision [6], Kangaroo [10], LongVU [23], and LongVA [31]. The performance of these models on video data is significantly worse compared to their performance on image data, with accuracy declining as the number of objects increases. This drop in accuracy can be attributed to specific challenges in video data, such as recurrence and scale variance, which are less prominent in images. Consequently, while the models approximate answers well, they often fail to deliver exact predictions, as illustrated in Fig. 4. Although commercial models perform somewhat better, their results remain comparable to the open-source state-of-the-art models.

GPT-4o demonstrates strong performance with low object counts (below 3), achieving over 80% accuracy, but its accuracy drops to below 50% as the object count increases. LLaVa-OneVision and LongVU show similar trends, and interestingly, both predict an object count of 8 more accurately than other mid-range object frequencies, indicating potential bias in their training data. LongVA frequently classifies instances into the ”ND” (Not Defined) category, reflecting uncertainty or misclassification. Kangaroo, which performs well in predicting up to 5 objects in image-based tasks, struggles significantly with counts below 3 in videos.

5.3. Case Study: Prompting with MCQs

Answering vs Estimating. Long VA consistently underperforms compared to other models as depicted in Fig. 3 and Fig. 4, often predicting ”undefined”. This suggests that Long VA is struggling to produce meaningful predictions, possibly due to a lack of clarity or difficulty in dealing with certain inputs, resulting in ambiguous or undefined outputs. To judge models approximation capabilities we modify prompt by providing multiple choices via augumenting the prompt with ”Choose the best answer from the following options.”

However, as observed in Fig. 7 even with the improvements observed in the multiple-choice (MCQ) scenario, Long VA still underperforms compared to better-performing open-source models like Long VU. This performance gap can be attributed to Long VA’s visual encoding strategy. Long VA represents videos as extended images using a unified encoding scheme, which limits its ability to handle dynamic temporal information effectively. The model’s emphasis on maintaining unlimited context length, while beneficial in theory, compromises its ability to capture the nuanced changes between frames in video data. As a result, Long VA struggles with tasks that require understanding temporal progression, leading to overall weaker performance in comparison to models like Long VU, which are better at handling both static and temporal visual data.

Accuracies across Object Types. Utilising VisCount images for evaluation, the data was classified into four classes: *people*, *animal*, *food*, and *objects* to analyze potential class bias. The resultant collection contained 322 records for objects, 23 for food items, 99 for people, and 42 for animals.

For the VisCount videos, grouping was done into three categories: *animal*, *people*, and *object*. The collection contained 48 instances of objects, 97 instances of animals, and 35 instances of people.Group-wise accuracies for the 5 models in our experimentation were assessed, as shown in Fig. 6. We observe, in the case of images, that LongVU and Kangaroo perform consistently throughout all the classes. LLaVa-OneVision performs slightly better in animals and almost consistently in all other classes.

6. Limitations

VisCount provides a valuable benchmark for evaluating object counting in low-density, well-separated scenarios . One key future scope is addressing object counting in medium to high-density environments, where challenges like object occlusion and vanishing points become prominent. In such scenes, objects often overlap or disappear into the background, making accurate counting more difficult. Extending VisCount to include higher-density data could better evaluate models’ robustness in these complex scenarios.

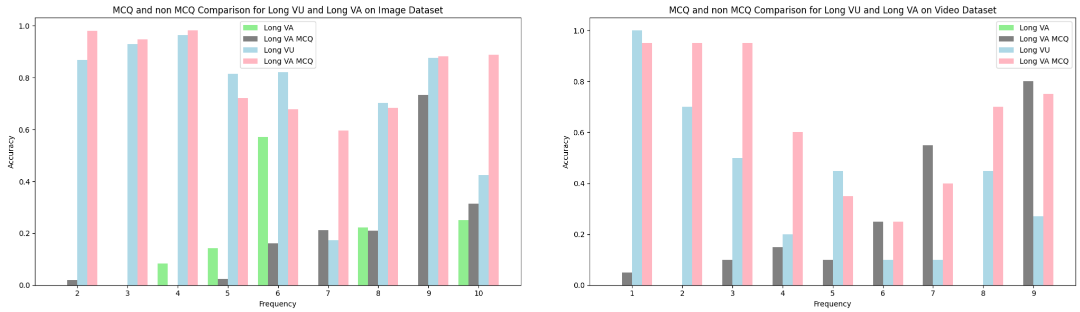


Figure 7. **Accuracy comparison with and without options on the proposed dataset VisCount** using prompt tuning of models. Evaluating estimating capabilities under-performing LongVA [31] with respect to a better performing SoTA LongVU [23]

The current video subset of VisCount contains only 20 examples per frequency class. Future work could involve expanding the number of video samples to provide a more comprehensive evaluation of object counting models in different temporal and environmental conditions.

While VisCount includes short video sequences (average of approximately 20 seconds), incorporating longer videos with more substantial background and scene transitions could help assess models' ability to handle dynamic environments with varying contexts. Finally, enhancing the dataset to capture long-range temporal dependencies would improve its utility for applications like surveillance or monitoring, where objects need to be tracked consistently over extended periods. Expanding VisCount in these directions would provide a more holistic benchmark, addressing a broader range of real-world object counting challenges.

7. Conclusion

We propose a novel dataset, VisCount, specifically designed to establish a benchmark for evaluating object-counting capabilities across varying object frequencies and classes, focusing on low-density and widely recognized objects. Our evaluations using VisCount on recent multimodal Video-LLMs, including the commercial GPT-4o [16] and open-source models LLaVa-OneVision [6], Kangaroo [10], LongVU [23], and LongVA [31], reveal significant limitations in their performance. Notably, these models perform worse on videos than on images, and their accuracy decreases as the frequency count of objects increases.

VisCount also sets the stage for addressing additional challenges, such as scale variation through scenes, object occlusions, and reappearances in videos, which are pivotal for robust video understanding. A wholesome dataset encompassing these scenarios would further evaluate models' capabilities in complex and dynamic environments, fostering advancements in the field. While multimodal Video-LLMs claim to possess strong video understanding capabilities, these results demonstrate their inability to accurately count objects— a fundamental task underpinning many aspects of visual comprehension, such as tracking, temporal reasoning, and scene analysis. This discrepancy highlights a critical gap in their ability to perform essential vision tasks, despite their broader claims of competence.

Our findings underscore the need for improved model architectures that can better handle temporal dynamics and adapt to increasing object densities. Additionally, there remains a critical gap in existing datasets for object counting, particularly in real-world scenarios with higher-density and dynamic environments. By addressing these limitations, VisCount provides a foundation for future research aimed at advancing multimodal Video-LLMs and their applications.

References

[1] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. Counting Everyday Objects in Everyday Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4428–4437, Honolulu, HI, 2017. IEEE. 1, 2

[2] Ana Cláudia Akemi Matsuki de Faria, Felype de Castro Bastos, José Victor Nogueira Alves da Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves dos Santos. Visual question answering: A survey on techniques and common trends in recent literature, 2023. 2

[3] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[4] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis, 2024. arXiv:2405.21075. 3

[5] Muhammad Uzair Khattak, Muhammad Ferjad Naeem,

Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. How Good is my Video LMM? Complex Video Reasoning and Robustness Evaluation Suite for Video-LMMs, 2024. arXiv:2405.03690 [cs]. 3

[6] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer, 2024. arXiv:2408.03326 [cs]. 1, 2, 4, 5, 6, 7, 8

[7] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 2

[8] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multimodal video understanding benchmark, 2024. 3

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1

[10] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A Powerful Video-Language Model Supporting Long-context Video Input, 2024. arXiv:2408.15542 [cs]. 2, 4, 5, 6, 7, 8

[11] Onalenna J Makhura and John C. Woods. Video Object Counting Dataset. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 1–4, 2019. 2, 3

[12] R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024. 2

[13] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. 2

[14] Melanie Mitchell. The LLM Reasoning Debate Heats Up, 2024. 2

[15] Thong Thanh Nguyen, Zhiyuan Hu, Xiaobao Wu, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. Encoding and controlling global semantics for long-form video question answering, 2024. 1, 2

[16] OpenAI. Gpt-4o release, 2024. 1, 4, 5, 6, 7, 8

[17] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to Count to Ten, 2023. arXiv:2302.12066 [cs]. 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. ISSN: 2640-3498. 2

[19] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark, 2024. 1, 2

[20] Michael S. Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Silvio Savarese, Ran Xu, Caiming Xiong, and Juan Carlos Niebles. xgen-mm-vid (blip-3-video): You only need 32 tokens to represent a video even in vlms, 2024. 2

[21] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs, 2021. arXiv:2111.02114 [cs]. 3

[22] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM International Conference on Multimedia*, Mountain View, CA USA, 2017. 4

[23] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding, 2024. arXiv:2410.17434 [cs]. 2, 4, 5, 6, 7, 8

[24] Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method, 2020. 1, 2

[25] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2024. 1, 2

[26] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible Object Counting in Underwater Scenes, 2023. arXiv:2304.11677 [cs]. 2

[27] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. *CoRR*, abs/2111.14690, 2021. 4

[28] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. 2

[29] Nan Xu and Xuezhe Ma. Llm the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems, 2024. 2

[30] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 1

[31] 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, 2024. arXiv:2406.16852 [cs]. 2, 4, 5, 6, 7, 8

[32] Yunhua Zhang, Ling Shao, and Cees G. M. Snoek. Repetitive Activity Counting by Sight and Sound, 2021. arXiv:2103.13096. 2

[33] 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, 2024. arXiv:2406.04264 [cs]. 3