ImageSequence : A Benchmark for Visual-Temporal Reasoning

Anonymous ACL submission

Abstract

Understanding how events occur over a time period is fundamental to human perception and reasoning, yet current Multimodal Large Language Models (MLLMs) largely treat vision as a static task, such as object recognition or caption generation rather than event progression. To bridge this gap, we introduce a benchmark designed to evaluate whether MLLMs can infer the correct temporal order of events from a set of images and a short textual description for context. Each instance in the benchmark contains a reference image, a textual description, and multiple unordered images depicting different stages of the same event. The model must rearrange these images to form a legit sequence. We evaluate some open-source MLLMs on the benchmark and the results reveal that even the strongest models, achieve moderate accuracy, indicating that existing MLLMs still lack a understanding of temporal order. These findings suggest that visual–temporal reasoning remains a major unsolved challenge for current multimodal models.

Github-Repo: `https://github.com/srajan0149/ImageSeq_Benchmark`
Website: `https://srajan0149.github.io/ImageSeq_Benchmark/`

1 Introduction

Models like GPT-4V (Achiam et al., 2023), BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023), and Flamingo (Alayrac et al., 2022) have integrated visual modality into large language models, enabling them to process and reason over images along with text. Many state-of-the-model MLLMs are still struggling with visual commonsense reasoning.

Given a set of related unordered images, humans can infer the order with ease, this ability to reason what comes next by inferring is central to how humans construct meaning, causality and establishing narrative logic. On the other hand, in Multimodal Large Language Models (MLLMs), this kind of reasoning is underexplored. Most of the MLLMs succeed at connecting visual and textual data in tasks such as captioning, visual question answering and retrieval, but rarely show in capabilities to deal with understanding of temporal understanding and causal dependency across the unordered image sequence.

While earlier models explored temporal order understanding, but the focus was primarily on better feature representation and learning rather than on reasoning. Similarly, earlier works on visual storytelling assumed pre-ordered image sequences and focused on text generation, leaving unordered sequence visual storytelling infused with temporal order understanding underexplored.

To address this gap, we introduce a benchmark to evaluate visual story ordering, helping model to infer the correct order of events.

2 Related Work

Earlier work on computer vision, explored around, teaching the model to understand sequence and progression. In *Shuffle and Learn* (Misra et al., 2016), model was trained to decide if the shuffled video frames shown are in the correct order or not. This achievement started the race for learning temporal features without labels.

With the introduction of Rank Pooling (Fernando et al., 2016), models learn to rank the frame-level features of a video in a chronological manner, which gives a new representation that captures the video-wide temporal understanding of a video, and its down streamed to tasks such as action recognition.

Similarly models were trained to understand the arrow of time (Wei et al., 2018), able to distinguish if the sequence is in forward or reverse direction, effectively grounding them in concept of temporal asymmetry. Along the same time visual storytelling (Huang et al., 2016; Feng et al., 2023) came into picture, but it implicitly assumed the order of images and did evaluate the ability to recover the order.

Many recent benchmarks examine the models temporal understanding ability but, it emphasizes on motion continuity rather than casual or narrative sequencing, leaving a wide gap to be filled in evaluation of MLLMs.

To bridge this gap, we introduce a benchmark designed to evaluate visual story ordering, capability of an MLLM from an unordered set of images in a casual or narrative sequencing manner.

3 Task Definition

The benchmark is designed to evaluate the model's ability to reason across relative temporal order between shuffled images using given textual context.

Each instance of the benchmark is represented as:

$$(I_r, T_r, \mathcal{I}_u) \quad \text{where} \quad \mathcal{I}_u = \{I_1, I_2, \ldots, I_k\} \quad (1)$$

- I_r: Is reference image representing the anchor event, based on which ordering will happen.
- T_r: Textual description providing the context for ordering.
- $\mathcal{I}_u$: Unordered set of related images showing the event at different .

We are evaluating the model ability to output an ordered sequence:

$$\hat{S} = [\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_k] \quad (2)$$

which is consistent with the temporal structure implied by the reference image and the textual description.

Evaluation We evaluate the model's ability to order the sequence using two metrics.

Pairwise Ordering: Given (I_r, T_r, I_a, I_b), the model predicts whether I_a occurs *before* or *after* I_b relative to the reference context. This tests local temporal reasoning and event understanding.

Sequence Ordering: Given the unordered set $\mathcal{I}_u$, the model outputs the ordered sequence $\hat{S}$. This tests global temporal understanding.

We used following computing resources for the benchmark source code:
300 GB of Disk space.
48 GB of VRAM (GPU RAM)
1 L40S GPU used

4 Dataset Construction

The benchmark is constructed to evaluate the MLLMs performance in temporal reasoning using real-world video frames with human textual context. Each instance in the dataset corresponds to a youtube video that contains a logical event, Annotators manually select a small set of key timestamps, ensuring that each frame picked depicts a logical flow of the event. The video is then processed using a custom pipeline and each extracted frame using the pipeline is uniquely labeled using the hashed YouTube ID and timestamp for consistency and reproducibility.

Annotators identify one frame as the reference image, which is used to arrange the unordered related images, and also provide a short textual description summarizing the overall activity in relevance to this reference image.The remaining frames from the same video is the candidate image set, containing unordered images of the event.

Together, each instance forms a tuple consisting of the reference image, the textual description, and the candidate set.

Currently the benchmark is divided into domains such as culture, daily routine, historical events, nature, and sports. Here is what each caegory means:

- **Culture:** Culture category contains image sequences of movies mainly of Hindi and English, and also of Tamil, Malayalam, Bengali and other Indian languages. Image sequences of weddings, celebrations, festivals and ceremonial events are also added.
- **Daily routine:** This category contains image sequences of 25 different activities like waking up, washing the car and getting a haircut.
- **Historical events:** This category contains 88 image sequences related to the history of places (eg. Greece and Japan), empires and activities like Cricket, Film production and cars.
- **Nature:** Nature category contains image sequences of 100 different natural phenomenon like blooming flowers, flow in water bodies and growth in different species.
- **Sports:** Sports category contains image sequences of 80 events of miscellaneous sports like Cricket, Football and Badminton.

5 Chosen Models for Evaluation

To assess the ability of current MLLMs to reason about temporal and casual order, we evaluate state of the art models that vary in scale, architecture, and pretraining strategy. The models are selected from the publicly available **OpenCompass VLMEvalKit** "Open VLM Leaderboard", hosted on Hugging Face.

The selected models include the InternVL (Chen et al., 2024), Ovis2 (Lu et al., 2024), Qwen (Bai et al., 2025), Kimi, and MiniCPM families, consisting of both large instruction-tuned models and lightweight efficient variants. Specifically, we test OpenGVLab's InternVL3-14B, InternVL3-8B, InternVL2.5-8B-MPO, and InternVL2.5-4B-MPO, AIDC-AI's Ovis2-16B, Ovis2-8B, and Ovis2-4B, Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Kimi-VL-A3B-Instruct (Team et al., 2025) and MiniCPM-o-2.6 (Yao et al., 2024).

The InternVL (Chen et al., 2024) series represents a strong open-source baseline designed for fine-grained multimodal alignment, while the Ovis2 (Lu et al., 2024) models emphasize robust instruction tuning and structured reasoning across modalities. Qwen2.5-VL (Bai et al., 2025) and Kimi-VL (Team et al., 2025) are general-purpose instruction-following MLLMs, whereas MiniCPM-o-2.6 (Yao et al., 2024) serves as a compact model being used to study scaling effects on temporal understanding.

The choice of evaluating on such diverse types of architectures, was to help us analyze how visual grounding, instruction tuning, and model capacity influence a model's ability to infer the task that we have defined. All models are assessed in a zero-shot setting using a common prompt that provides a reference image, a textual description, and a set of unordered images.

The task requires models to output the most probable chronological sequence, capturing both pairwise order and global sequence order. Performance is evaluated using pairwise and sequence accuracy, telling us how well the models can perform the task.

6 Result And Analysis

The table 1, reports the pairwise across all the defined domains of the proposed benchmarks for five models, InternVL3-8B (Chen et al., 2024), Ovis2-16B (Lu et al., 2024), Ovis2-8B (Lu et al., 2024), Ovis2-4B (Lu et al., 2024), and Qwen2.5-VL-7B (Bai et al., 2025).

These models are evaluated on the benchmark using pairwise accuracy, and the results are reported as mean and median accuracy for each domain. The performance across all domain for all models remain in the range of 0.44 to 0.50, indicating that exisiting MLLMs model struggle with the task defined.

Among the evaluated models, Qwen2.5-VL-7B (Bai et al., 2025) achieved the highest overall mean accuracy, followed by Ovis2-16B (Lu et al., 2024) and Ovis2-4B (Lu et al., 2024). Across domains, the Daily Routine category consistently showed the highest accuracy for all models with mean accuracy ranging between 0.60 to 0., suggesting that models perform better when events look familiar, everyday sequences that align with their pretraining data. In contrast, Nature and Sports categories recorded the lowest accuracy, ranging from 0.40 to 0.46. For three image sequences of Nature category, the output was invalid.

We also observe that the model size does not guarantee improved temporal reasoning, the smaller Ovis2-4B (Lu et al., 2024) performed comparably to its 8B and 16B counterparts.

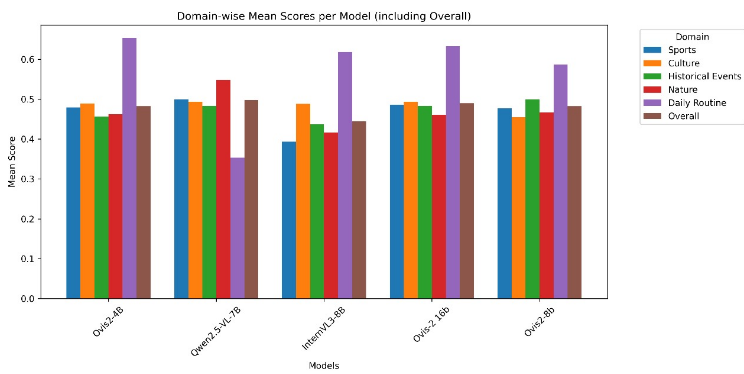

Figure 1: Domain Wise Mean Scores Per Model

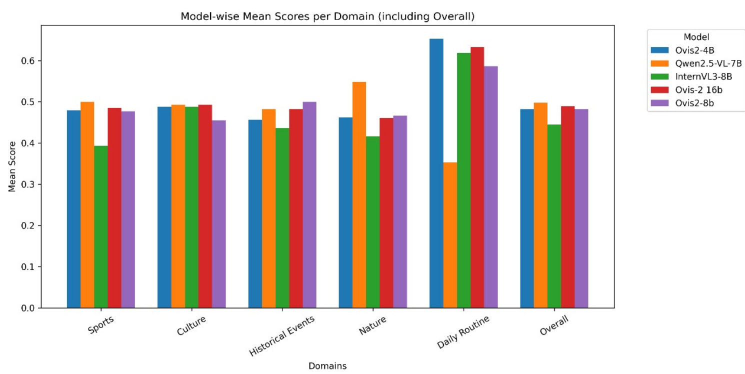

Figure 2: Model Wise Mean Scores Per Domain

Table 1: Pairwise accuracy across domains for different MLLMs on the proposed benchmark.

Model	Domain	Total	Mean Acc.	Median Acc.
InternVL3-8B	Culture	74	0.4883	0.5000
	Daily Routine	25	0.6187	0.6667
	Historical Events	88	0.4371	0.5000
	Nature	97	0.4158	0.3333
	Sports	80	0.3933	0.4000
	Overall	364	0.4447	0.4000
Ovis2-16B	Culture	74	0.4932	0.5000
	Daily Routine	25	0.6333	0.6667
	Historical Events	88	0.4830	0.5000
	Nature	100	0.4610	0.5000
	Sports	80	0.4854	0.5000
	Overall	367	0.4898	0.5000
Ovis2-8B	Culture	74	0.4550	0.5000
	Daily Routine	25	0.5867	0.6667
	Historical Events	88	0.5000	0.5000
	Nature	100	0.4667	0.5000
	Sports	80	0.4771	0.5000
	Overall	367	0.4827	0.5000
Ovis2-4B	Culture	74	0.4887	0.5000
	Daily Routine	25	0.6533	0.6667
	Historical Events	88	0.4564	0.5000
	Nature	100	0.4617	0.5000
	Sports	80	0.4792	0.5000
	Overall	367	0.4827	0.5000
Qwen2.5-VL-7B	Culture	74	0.4932	0.5000
	Daily Routine	25	0.3533	0.3333
	Historical Events	88	0.4830	0.5000
	Nature	100	0.5483	0.6667
	Sports	80	0.5000	0.3333
	Overall	367	0.4977	0.5000

7 Future Work And Conclusion

While our benchmark evaluates the model on temporal relation between unordered set of images. The current dataset focuses only on short and simple events, we plan to extend it to include a large number of image sequence, covering complex and longer events with complex scenes and multiple actors. We also plan to test proprietary MLLMs to evaluate, how closed-source models perform compared to open-sourced models.Beyond explanation , tasks such as next image generation can also be explored, where the model would have to synthesis the next likely frame in the sequence given the textual context.

8 Feedback

The feedback noted here is based on viva of assignments 1 and 2 :

- Less reference to the existing benchmarks mentioned in the paper
- Did not provide the working principle behind the metrics used, such as ROUGE, BLEU, bertscore and chrF++.
- Did not incorporate reasoning for the patterns in the results.
- Missing contribution section in the Assignment 2 report.

- Need to look into grammatical mistakes

We have tried our best to incorporate the feedback that has been given to us.

9 Contributions

- Naren Kumar S: Literature review and paper writing
- Rahul Khichar: Metric calculations
- Srajan Dehariya: Programming and execution
- Dhruv Goel: Website development
- Simran: Report writing
- Naren Kumar S, Rahul Khichar, Srajan Dehariya, Dhruv Goel, Jeet Joshi, Simran, Pranav Somase, Sai Krishna: Ideation and Dataset curation

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zhangyin Feng, Yuchen Ren, Xinmiao Yu, Xiaocheng Feng, Duyu Tang, Shuming Shi, and Bing Qin. 2023. Improved visual story generation with adaptive context modeling. *arXiv preprint arXiv:2305.16811*.

Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. 2016. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, and 1 others. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *European conference on computer vision*, pages 527–544. Springer.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 73 others. 2025. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. 2018. Learning and using the arrow of time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8052–8060.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.