MVTamperBench: Evaluating Robustness of Vision-Language Models

Amit Agarwal[1*+], Srikant Panda[2*], Angeline Charles[3*], Hitesh Patel[5] Bhargava Kumar[4], Priyaranjan Pattnayak[6], Taki Hasan Rafi[7], Tejaswini Kumar[4], Hansa Meghwani[1], Karan Gupta[5], Dong-Kyu Chae[7+]

[1]Liverpool John Moores University [2]Birla Institute of Technology [3]Christ University
[4]Columbia University [5]New York University [6]University of Washington [7]Hanyang University
[*]Equal Contribution [+]Corresponding Authors

Correspondence: amit.pinaki@gmail.com, dongkyu@hanyang.ac.kr

Abstract

Multimodal Large Language Models (MLLMs), are recent advancement of Vision-Language Models (VLMs) that have driven major advances in video understanding. However, their vulnerability to adversarial tampering and manipulations remains underexplored. To address this gap, we introduce **MVTamperBench**, a benchmark that systematically evaluates MLLM robustness against five prevalent tampering techniques: rotation, masking, substitution, repetition, and dropping; based on real-world visual tampering scenarios such as surveillance interference, social media content edits, and misinformation injection. MVTamperBench comprises 3.4K original videos, expanded into over 17K tampered clips covering 19 distinct video manipulation tasks. This benchmark challenges models to detect manipulations in spatial and temporal coherence. We evaluate 45 recent MLLMs from 15+ model families. We reveal substantial variability in resilience across tampering types and show that larger parameter counts do not necessarily guarantee robustness. MVTamperBench sets a new benchmark for developing tamper-resilient MLLM in safety-critical applications, including detecting clickbait, preventing harmful content distribution, and enforcing policies on media platforms. We release all code, data, and benchmark to foster open research in trustworthy video understanding.

1 Introduction

Multimodal Large Language Models (MLLMs) have catalyzed significant progress in video understanding, enabling a wide array of applications across domains such as surveillance, healthcare, and autonomous systems. However, their growing integration into high-traffic platforms (e.g., Instagram, Facebook, TikTok) has exposed critical vulnerabilities. Specifically, tampered videos are increasingly exploited to bypass platform policies, disseminate harmful content, and promote clickbait, posing serious challenges for content moderation and policy enforcement (Kingra et al., 2023; Times of India, 2024).

These threats underscore the urgent need to improve the robustness of MLLMs against real-world manipulations. Although video tampering can involve audio modifications, synthetic speech overlays, or deepfake generation, our benchmark focuses exclusively on visual-only manipulations. This choice is driven by the current limitations of existing models, most of which currently lack comprehensive audiovisual processing capabilities.

Unlike adversarial robustness in static images, an area that has been extensively studied, video tampering introduces unique challenges that arise from the interplay of spatial and temporal dynamics. Common manipulation techniques such as frame dropping, masking, repetition, substitution, and rotation disrupt this coherence, frequently resulting in catastrophic model failures. These methods mirror real-world adversarial tactics: substitution injects objectionable material (e.g., nudity, violence) to circumvent detection; dropping eliminates key surveillance evidence; repetition loops footage to obscure illicit activity; masking occludes critical regions; and rotation induces spatial distortions commonly seen in edited or re-uploaded content.

Existing adversarial and traditional approaches, including black-box (Jiang et al., 2019) and cross-modal (Wei et al., 2021) attacks, primarily address isolated scenarios rather than systematically evaluating diverse tampering types. Furthermore, existing multimodal benchmarks (Pattnayak et al., 2024; Agarwal, 2021)—such as MMBench-Video (Fang et al., 2024), BLINK (Fu et al., 2025), and Video-MME (Fu et al., 2024)—focus on multimodal comprehension and temporal reasoning but overlook *adversarial* robustness. For example, MMBench-Video evaluates cross-modal alignment

Findings of the Association for Computational Linguistics: ACL 2025, pages 18804–18828

without adversarial testing; BLINK targets long-form temporal reasoning without addressing tampering impacts; and Video-MME, while effective for vision-language alignment, omits tampering-specific tasks. This leaves a critical gap in systematically assessing how MLLMs withstand real-world manipulations.

To bridge this gap, we introduce **MVTamperBench**, a benchmark specifically designed to evaluate MLLM robustness against five prevalent tampering techniques. Built from 3.4K original videos—expanded into over 17K tampered clips spanning 19 video tasks—MVTamperBench challenges models to detect manipulations by disrupting temporal and spatial coherence in a video. By focusing on tampering resilience, we believe that MVTamperBench provides a critical tool for improving MLLM robustness, and eventually, it enables the development of tamper-resilient models applicable to real-world challenges such as clickbait detection, content moderation, and policy enforcement, advancing adversarial robustness in video understanding for high-stakes domains.

Our main contributions are as follows:

- We introduce **MVTamperBench**, a benchmark that systematically evaluates MLLMs on five major video manipulations, focusing on spatial and temporal coherence to stimulate real-world scenarios.
- We propose a unified evaluation methodology that frames tampering detection as a multiple-choice task, enabling straightforward, interpretable and consistent performance comparisons.
- Through experiments on 45 MLLMs across 15+ families, we identify critical vulnerabilities across MLLM families. across MLLM families, without any correlation between model size and performance.
- Our released code enables researchers to integrate additional datasets and adapt or add new tampering, facilitating domain-specific extensions, and supports reproducibility.

2 Related Work

The development of benchmarks for MLLMs has significantly advanced the evaluation of image and video understanding tasks. They have covered spatial reasoning, temporal comprehension, object detection, common sense inference, and so on. However, the robustness of MLLMs to adversarial manipulations, particularly video tampering, remains underexplored. This section reviews existing benchmarks, which can be categorized into image-based and video-based evaluations, and analyzes their contributions and shared limitations.

2.1 Image-based Understanding

Benchmarks like BLINK (Fu et al., 2025) and MuirBench (Wang et al., 2024a) focus on evaluating static visual reasoning. BLINK tests foundational tasks such as depth estimation, forensic detection, and visual correspondence, which cover a diverse set of challenges for spatial reasoning. Similarly, MuirBench extends this evaluation to multi-image tasks, including action recognition and geographic reasoning, by synthesizing information from diverse sources. While these benchmarks have advanced static image understanding, their reliance on single or multiple still images excludes temporal dynamics and limits their applicability to scenarios involving sequential manipulations, such as those found in video content.

2.2 Video-based Understanding

Video-based benchmarks have expanded the scope of evaluation by incorporating temporal and multimodal reasoning. MVBench (Li et al., 2024e), MMBench-Video (Fang et al., 2024), and Video-MME (Fu et al., 2024) focus on tasks such as event detection, episodic reasoning, and contextual understanding. These benchmarks challenge models with diverse tasks spanning object interactions, long-duration video analysis, and domain-specific reasoning. Similarly, LongVU (Shen et al., 2024) introduces spatio-temporal compression techniques to enhance efficiency, while MotionEpic (Fei et al., 2024a) integrates Spatial-Temporal Scene Graphs (STSG) for fine-grained cognitive tasks. However, while these benchmarks assess temporal coherence and reasoning, they do not systematically address adversarial robustness, particularly in tampering scenarios.

Other benchmarks, such as Wolf (Li et al., 2024a) and Sharingan (Chen et al., 2024a), target specialized video understanding tasks. Wolf focuses on captioning using a mixture-of-experts strategy, while Sharingan extracts action sequences from desktop recordings using frame-differential approaches. Although these benchmarks achieve

Figure 1: Illustration of the five video frame tampering techniques. Each row shows a specific tamper type applied to a video snippet, describing the respective impact on temporal/spatial coherence.

high accuracy in their domains, they are limited to specific tasks and lack general mechanisms to evaluate the resilience to tampering.

In summary, despite their contributions to advancing MLLM evaluation, existing benchmarks largely focus on performance under ideal conditions and neglect robustness to tampering or adversarial effects. Techniques such as frame substitution, masking, repetition, dropping, and rotation disrupt temporal coherence and pose unique challenges for multimodal models. The absence of systematic evaluations of these tampering techniques highlights a critical gap in ensuring model reliability for real-world applications like forensic analysis, media verification, and misinformation detection.

3 MVTamperBench

We introduce **MVTamperBench**, a comprehensive benchmark designed to systematically evaluate the robustness of MLLMs against video tampering techniques. Through our MVTamperBench which introduces diverse manipulations, we aim to broaden the evaluation landscape, enabling a deeper understanding of model strengths and vulnerabilities under adversarial scenarios. Table 3 compares MVTamperBench with existing benchmarks, highlighting its focus areas, strengths, and unique contributions.

In the following subsections, we detail its construction, design choices, and key features. More details can be found in our code[1], data[2] and benchmark [3] repositories.

3.1 Benchmark Construction

To evaluate MLLM robustness under adversarial conditions, we apply distinct tampering methods as shown in Figure 1 —*Dropping*, *Masking*, *Substitution*, *Repetition*, and *Rotation*—to the 3,487 original MVBench (Li et al., 2024e) videos (excluding NTU dataset due to licensing), resulting in a total of 17,435 tampered clips. These manipulations target both spatial and temporal coherence, thereby simulating common real-world tampering scenarios such as deliberate frame editing or slicing from unrelated content.

[1]https://amitbcp.github.io/MVTamperBench/
[2]https://hf.co/datasets/Srikant86/MVTamperBench
[3]https://github.com/open-compass/VLMEvalKit

3.2 Tampering Techniques

Dropping: Removes a 1-second segment for creating temporal discontinuity.
Masking: Overlays a black rectangle on a 1-second segment. It aims to simulate visual data loss.
Rotation: Rotates a 1-second segment by 180 degrees for introducing spatial distortion.
Substitution: Replaces a 1-second segment with a pre-selected clip from another video, in order to disrupt temporal and contextual flow.
Repetition: Repeats a 1-second segment, introducing temporal redundancy.
The aforementioned effects are applied uniformly across all videos to ensure consistent and comparable evaluation.

3.3 Design & Implementation

Tampering Duration (1 Second). We fix tampering to a 1-second segment to align with reports of minimal but impactful real-world tampering, e.g., short edits on social networks or subtle modifications on surveillance feeds (Alicia, 2024). Our preliminary experiments revealed that using less than 1 second could be overlooked by certain model sampling mechanisms, whereas longer tampering (>3s) often resembled normal scene transitions, reducing adversarial impact.

Tampering Location (Middle). All manipulations occur at the video's midpoint to disrupt central content. Our pilot tests showed that tampering near the start or end risked mimicking scene cuts or information loss, which makes detection less indicative of genuine adversarial robustness.

Substitution Source. For Substitution, the 1-second clip is randomly chosen from a consistent pool of different videos within MVBench. This ensures uniform difficulty across all samples, thus preventing confounds from domain shifts or overly simplistic substitutes.

Modular & Scalable Framework. Each technique is encapsulated in a reusable class for facilitating custom parameterization (e.g., tampering duration, location, intensity). Our open-source code will integrate seamlessly with **VLMEvalKit** (Duan et al., 2024), thereby promoting reproducible experiments and easy extension to other datasets or tasks.

To inform our final design decisions, we conducted a series of exploratory experiments examining the effects of tampering position and duration. Results from these alternative design configurations are detailed in Appendix A.2.1.

3.4 Dataset Scope and Statistics

All 3,487 MVBench videos undergo each of the five tampering types, producing 17,435 tampered clips. The five manipulations are uniformly applied to ensure comparability across different MLLM architectures and training regimes. Figure 2 illustrates the distribution of their durations. We can observed that diverse scenarios from short (3-5s) to extended (>20s) sequences are included.

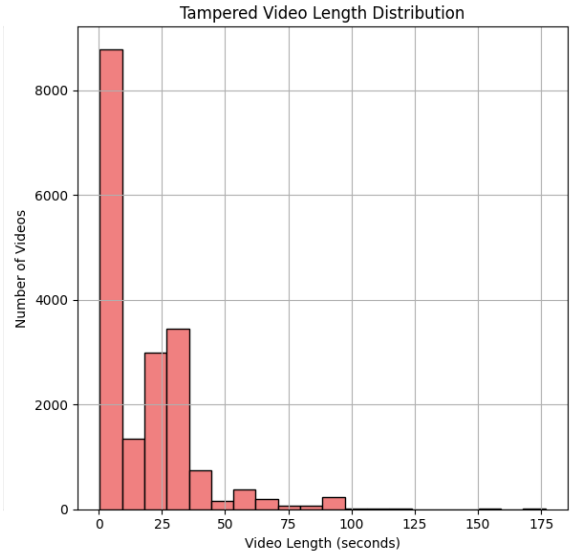


Figure 2: Distribution of video durations. Our dataset spans a broad range of durations, which can reflect varied real-world conditions.

3.5 Summary

By enforcing consistent parameters (1s duration, midpoint placement) and systematically applying five tampering methods, **MVTamperBench** offers a controlled yet flexible platform for evaluating tampering resilience. Our design can be easily extended with additional manipulations (e.g., noise injection, partial masking, positions), deepfakes or integrated into domain-specific contexts like surveillance feeds analysis or clickbait detection. We provide additional details on MVTamperBench, including video sources and associated tasks, in Appendix A.1.

4 Experiments

This section outlines the experimental setup, results, and analysis of 45 models evaluated on our proposed benchmark. Our results highlight their

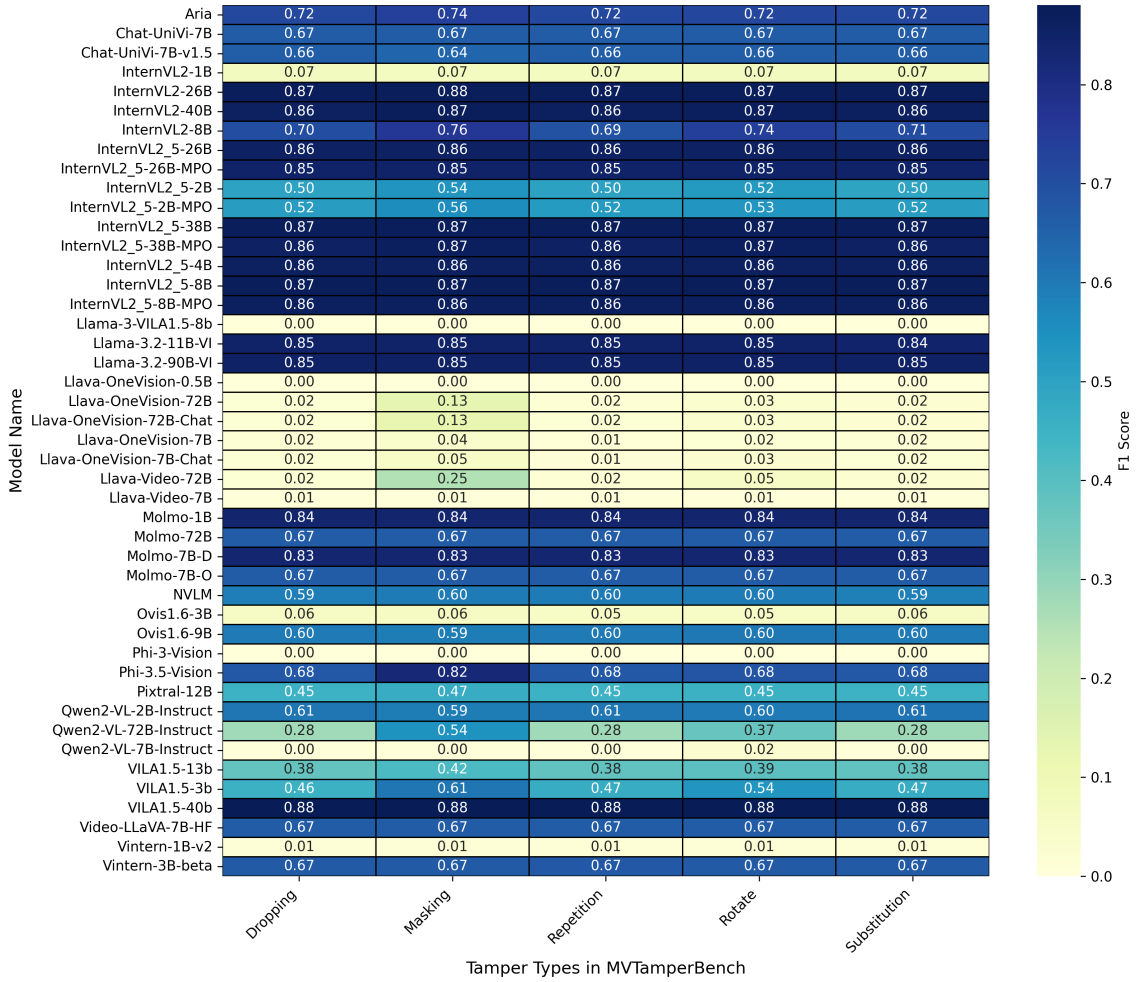


Figure 3: F1 scores across models and tampering types. High-performing models are robust across all types.

strengths, weaknesses, and actionable insights. We provide further overview of the evaluated MLLMs in Appendix A.3.

4.1 Experimental Setups

Evaluation Protocol Each model is tasked with identifying whether a video has been tampered with or not. For every video, the model is presented with the following structured prompt:

> *Does this video exhibit any signs of tampering, such as corruption, black-outs, rotated frames, repeated frames, or swapped frames?*
>
> **Options:** A. Yes B. No

The dataset comprises both tampered and non-tampered videos. For each tampered video, the corresponding non-tampered video is included to ensure a balanced distribution. Models must select one of the two options (**Yes** or **No**), and their predictions are compared against ground truth labels to determine correctness. This task is repeated for all tampering types.

We also compared structured, general, and chain-of-thought prompts (Appendix A.2.2), finding that general and CoT variants led to higher false positives due to limited temporal reasoning in current MLLMs.

Metrics The primary evaluation metric is the **F1 Score**, chosen for its ability to balance precision and recall, particularly in scenarios where misclassifications (false positives and false negatives) can significantly impact robustness evaluation. We believe that this metric is the most suitable for our setup, where models must not only detect manipulations but also avoid false positives on non-tampered videos. To capture performance across all tampering types, we compute individual F1 scores for each tampering type. In addition, we also measure **F1 (overall) score** as the macro average of these individual F1 scores. We thus highlight per-tamper-type strengths and weaknesses as well as overall model performance.

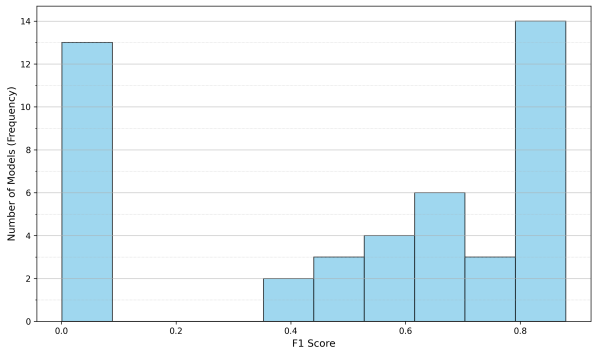


Figure 4: Distribution of F1 (overall) scores across models.

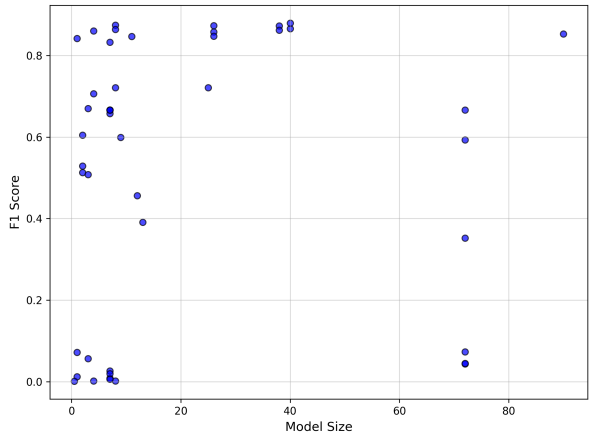


Figure 5: Scatter plot showing no correlation between model size and overall F1 (Pearson r=0.05).

4.2 Results and Analysis

We evaluate the performance of 45 MLLMs using F1 (overall) scores and individual F1 scores across five tampering types. The distribution of F1 (overall) scores (Figure 4) reveals significant variability in the robustness of the model, with several models struggling to detect tampering (F1 < 0.2), while a few high-performing models achieve F1 > 0.8.

We did not observe a correlation between model size and tampering detection performance (Pearson correlation = 0.05, Figure 5). This highlights that architectural differences and training techniques, rather than parameter count, contribute more significantly to robustness against tampering.

In addition, Figure 3 showcases model-specific adaptability across different tampering types. Models with F1 (overall) > 0.8 are consistently robust across all tampering types, while those with F1 < 0.2 perform slightly better on *Masking*, which relies more on spatial reasoning. Models in between generally struggle with temporal disruptions like *Dropping* and *Substitution*, reflecting challenges in temporal coherence.

4.2.1 Analysis based on Performance

Figure 6 categorizes MLLMs into **low-**, **moderate-**, and **high-** performing groups based on their F1-score distribution. We define the boundaries using the 0.25 quantile (F1 = 0.071) and 0.75 quantile (F1 = 0.846). We round-off the quantile thresholds to the nearest integers to identify low-performing models (F1 (overall) < 0.01) and high-performing models (F1 (overall) > 0.8).

Figure 7 illustrates the average F1 scores for each category. We observe stark differences between groups: High-performing models maintain consistent performance across all tampering types, whereas low-performing models show significant weaknesses, particularly for tampering effects that disrupt temporal coherence (*Dropping*, *Substitution*). Moderate-performing models excel slightly in *Masking* (Figure 8).

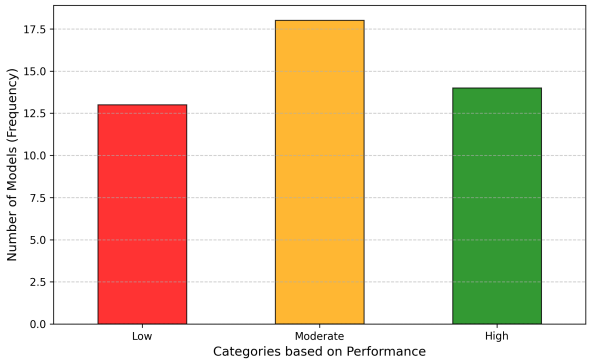


Figure 6: Distribution of Number of Models in low, moderate, and high-performing categories based on F1 (overall).

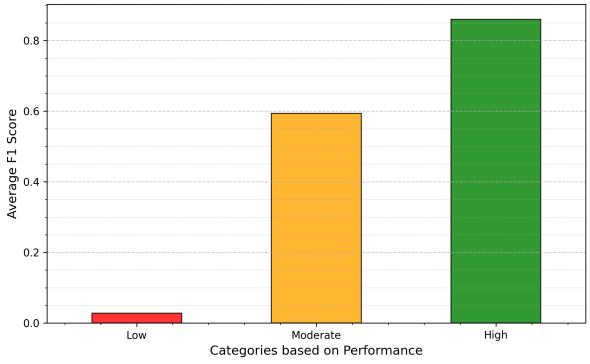


Figure 7: Average F1 (overall) scores across low, moderate, and high-performing model categories.

We analyze the variance in model performance between tampering types in Figure 9. High-performing models exhibit negligible variance, indicating consistent robustness across all tampering

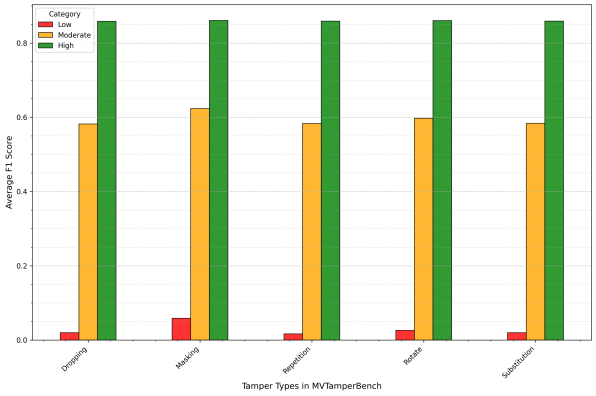


Figure 8: F1 scores for tampering types across model categories. Low & Moderate-performing models perform slightly better on *Masking*, while high-performing models show consistent robustness.

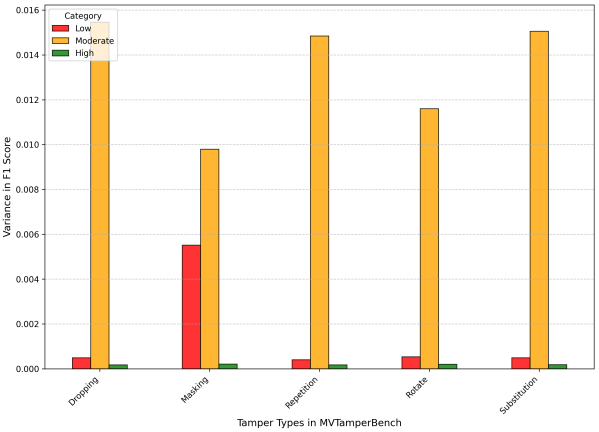


Figure 9: Variance in F1 scores across tampering types for each model category.

effects. This stability suggests that these models are well-equipped to handle both spatial & temporal disruptions introduced by tampering.

Low-performing models display minimal variance across most tampering types, with the exception of *Masking*, which shows a slightly higher variance. This highlights a specific weakness in processing visual obfuscations, likely due to their reliance on static features rather than robust temporal reasoning. The minimal variance across other tampering types suggests that these models fail uniformly, regardless of the type of manipulation.

Moderate-performing models demonstrate the **highest variance overall**, particularly for tampering types such as *Dropping*, *Repetition*, and *Substitution*. This behavior indicates an inconsistency in their ability to adapt to tampering effects. While these models achieve a balanced performance across easier tampering types, temporal disruptions like *Dropping*, *Repetition*, and *Substitution* pose significant challenges, leading to occasional spikes in variance. This suggests that moderate-performing models, while more capable than low-performing counterparts, still struggle to maintain robustness across a diverse set of tampering scenarios. We provide a detailed analysis of individual model performance in Appendix A.4.1, and present qualitative trends based on model architecture and training paradigms across all models studied in Appendix A.4.2, offering insights into the factors influencing performance.

4.2.2 Analysis based on Model Size

We categorize models based on their parameter sizes into **small** (<7B), **medium** (7B–26B), and **large** (>26B) groups (Figure 11). While Figure 12 shows that larger models generally achieve higher F1 (overall) scores, Figure 5 confirms no significant correlation between model size and tampering detection performance (Pearson correlation = 0.05). A closer examination of individual model trends across size categories reveals several noteworthy patterns discussed in Appendix A.4.3.

Trends Across Families. Across size categories, we observe distinct trends in model performance. The **VILA model family** consistently improves with size, with the exception of **VILA1.5-8B**, highlighting the scalability of its architecture for tampering detection. Similarly, **Qwen2-VL** demonstrates significant gains with increased parameters, though it trails behind other families in absolute performance.

The **Llama3.2-Vision** family, despite its scaling efforts, reveals the diminishing returns of increasing model size without architectural or training advancements. Meanwhile, **Molmo** models illustrate the importance of efficient design, as **Molmo-1B** outperforms its larger variants like **Molmo-72B**. Finally, the dominance of **InternVL-2.5** across all sizes highlights the benefits of balanced architecture & task-specific training strategies.

4.2.3 Analysis across Video Task Types

MVTamperBench comprises 19 video task categories, each evaluated across five tampering techniques. Figure 10 highlights the F1 (overall) scores for each task, averaged across all tampering types and models. While certain tasks exhibit high detection F1 score, others remain significantly more challenging. Appendix A.4.4 provides additional insights into the performance trends of each tampering type across different tasks.

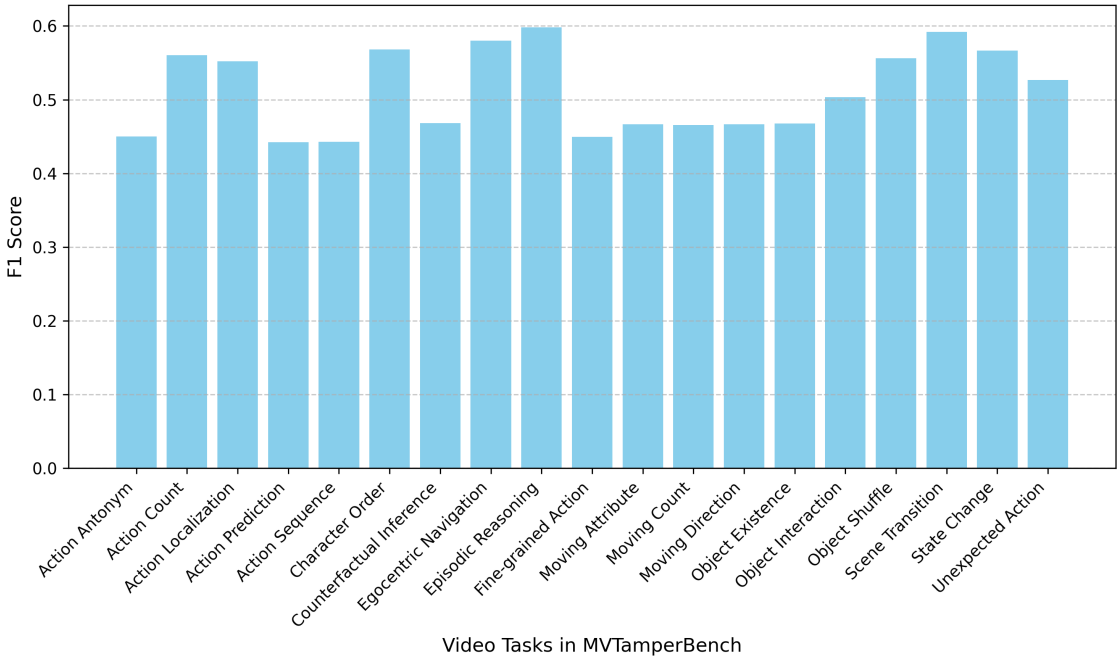


Figure 10: F1 (overall) scores for task categories in MVTamperBench. Tasks like *Episodic Reasoning* achieve higher scores, while *Counterfactual Inference* is more challenging.

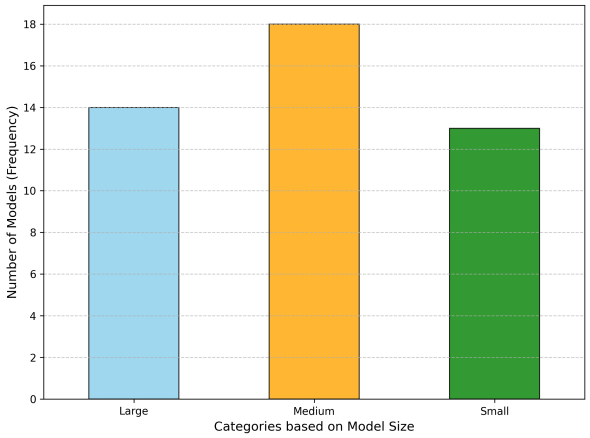


Figure 11: Distribution of models grouped by size; Categories: Small (<7B), Medium (7B–26B), and Large (>26B).

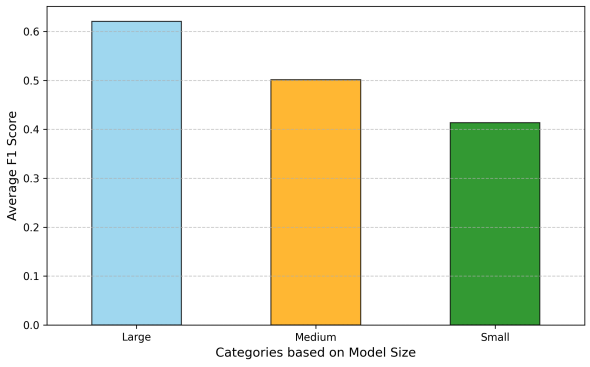


Figure 12: F1 (overall) scores for models by size category. Larger models generally tend to perform better.

Easier Tasks. Tasks such as **Episodic Reasoning, Scene Transition, Ego-centric Navigation,** and **State Change** consistently achieve higher F1 scores, as shown in Figure 10. These tasks often involve shorter temporal dependencies and simpler spatial reasoning, allowing models to rely on pre-trained vision-language features rather than advanced temporal reasoning.

For example, in **Episodic Reasoning**, tampering effects like *Dropping* and *Repetition* minimally affect task performance, as models primarily focus on memorizing what it has seen. Similarly, **Scene Transitions** tasks depend more on detecting static or localized changes, making them less susceptible to temporal disruptions such as *Dropping* or *Repetition*. The results indicate that models trained on extensive image-based datasets excel in tasks with lower temporal complexity.

Challenging Tasks. Tasks such as **Action Prediction, Counterfactual Inference,** and **Fine-Grained Action** pose significant challenges for tampering detection. These tasks inherently require complex temporal reasoning and context preservation, making them highly sensitive to disruptions introduced by tampering techniques.

In **Counterfactual Inference**, tampering effects like *Substitution & Dropping* cause significant confusion, as they disrupt the continuity of the video narrative required for hypothetical reasoning. Similarly, **Fine-Grained Action** detection is heavily

impacted by *Rotation*, distorting spatial relationships crucial for identifying subtle movements.

Interestingly, **Action Prediction** tasks also highlight the limitations of current MLLMs in understanding and reasoning about temporal progression. Models often fail to account for disruptions in video sequences, leading to degraded performance across tampering types.

We further discuss our Key Findings & Future Directions in Appendix A.6 & A.7 respectively.

5 Concluding Remarks

Novel Benchmark. We introduced MVTamperBench, a comprehensive benchmark for evaluating the robustness of Multimodal Large Language Models (MLLMs) against five key video tampering techniques—*Dropping*, *Masking*, *Repetition*, *Rotation*, and *Substitution*. Through systematic experiments on 19 video tasks involving 45 models across 15+ families, we observed pronounced variability in resilience. Notably, even MLLMs exceeding 70B parameters suffer severe performance drops, whereas select small models ($<$ 7B) demonstrate unexpectedly strong tampering detection, illustrating that size alone does not ensure robustness.

Impact on Research Community. Beyond revealing these vulnerabilities, we believe that MVTamperBench offers actionable insights for refining model architectures and training pipelines, highlighting gaps that must be addressed before these systems can be reliably deployed in safety-critical settings. Our open-source framework will support reproducible evaluations and community-driven extensions, enabling researchers to integrate additional datasets or adapt tampering methods for domains like clickbait detection, content moderation, and surveillance feeds analysis.

Future Work. Looking ahead, we plan to expand MVTamperBench with new tampering types (e.g., noise injection, frame shuffling) and domain-specific scenarios (e.g., healthcare, surveillance). By illuminating the nuanced impacts of video manipulations and guiding innovation in robust MLLM architectures, MVTamperBench lays a strong foundation for next-generation multimodal models capable of withstanding adversarial manipulations.

6 Limitation

MVTamperBench provides a robust framework for evaluating MLLM resilience against video tampering, but there are limitations that present opportunities for future work.

First, while the benchmark evaluates five tampering techniques, expanding to additional manipulations/modalities to better capture subtle, and emerging techniques. Second,the dataset currently relies on limited datasets. Expanding the benchmark to incorporate videos from diverse sources, such as user-generated content, surveillance footage, or policy enforcement, would broaden its applicability and relevance to task-specific and domain-specific challenges. Third, while the current evaluation focuses on binary detection, future benchmarks could assess a model's ability to classify the specific tampering type, providing deeper insights into its robustness. Finally, scalability to closed-source and extremely large-scale models ($>$ 100B parameters) remains a challenge due to computational and cost constraints.

Addressing these limitations will enable **MVTamperBench** to further advance tampering detection and resilience in a broader range of applications.

Acknowledgement

This work was partly supported by (1) the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00345398) and (2) the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University)).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Amit Agarwal. 2021. Evaluate generalisazion & robustness of visual features from images to video. *ResearchGate*. Available at https://doi.org/10.13140/RG.2.2.33887.53928.

Amit Agarwal and Kulbhushan Pachauri. 2023. Pseudo

labelling for key-value extraction from documents. US Patent 11,823,478.

Amit Agarwal, Kulbhushan Pachauri, Iman Zadeh, and Jun Qian. 2024a. Techniques for graph data structure augmentation. US Patent 11,989,964.

Amit Agarwal, Srikant Panda, Deepak Karmakar, and Kulbhushan Pachauri. 2024b. Domain adapting graph networks for visually rich documents. US Patent App. 18/240,480.

Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2024c. Synthetic document generation pipeline for training artificial intelligence models. US Patent App. 17/994,712.

Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2025a. FS-DAG: Few shot domain adapting graph networks for visually rich document understanding. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 100–114, Abu Dhabi, UAE. Association for Computational Linguistics.

Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2025b. Techniques of information extraction for selection marks. US Patent App. 18/240,343.

Amit Agarwal, Hitesh Patel, Priyaranjan Pattnayak, Srikant Panda, Bhargava Kumar, and Tejaswini Kumar. 2024d. Enhancing document ai data generation through graph-based synthetic layouts. *arXiv preprint arXiv:2412.03590*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Alicia. 2024. Is it illegal to tamper with security cameras? Accessed: 2025-01-11.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, et al. 2025. Crowdsource, crawl, or generate? creating sea-vl, a multicultural vision-language dataset for southeast asia. *arXiv preprint arXiv:2503.07920*.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. 2024. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*.

Y. Chen, Y. Ren, X. Qin, J. Zhang, K. Yuan, L. Han, Q. Lin, D. Zhang, S. Rajmohan, and Q. Zhang. 2024a. Sharingan: Extract user action sequence from desktop recordings. *arXiv preprint*, arXiv:2411.08768.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024d. 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.

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

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Khang T. Doan, Bao G. Huynh, Dung T. Hoang, Thuc D. Pham, Nhat H. Pham, Quan T. M. Nguyen, Bang Q. Vo, and Suong N. Hoang. 2024. Vintern-1b: An efficient multimodal large language model for vietnamese. *Preprint*, arXiv:2408.12480.

Karan Dua, Praneet Pabolu, and Mengqing Guo. 2024. Generating templates for use in synthetic document generation processes. US Patent App. 18/295,765.

Karan Dua, Praneet Pabolu, and Ranjeet Kumar Gupta. 2025. Generation of synthetic doctor-patient conversations. US Patent App. 18/495,966.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Amit Agarwal, Zhe Chen, Mo Li, Yubo Ma, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*.

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

X. Fang, K. Mao, H. Duan, X. Zhao, Y. Li, D. Lin, and K. Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*.

H. Fei, S. Wu, W. Ji, H. Zhang, M. Zhang, M. L. Lee, and W. Hsu. 2024a. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024b. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*.

Allen Institute for AI. 2024. Charades: A dataset for video understanding. https://prior.allenai.org/projects/charades. Accessed: 2024-10-28.

C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, and P. Chen. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W. C. Ma, and R. Krishna. 2025. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166, Cham. Springer.

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

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. *arXiv preprint arXiv:2501.02955*.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*.

Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. 2019. Black-box adversarial attacks on video recognition models. *Proceedings of the 27th ACM International Conference on Multimedia*.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13700–13710.

Staffy Kingra, Naveen Aggarwal, and Nirmal Kaur. 2023. Emergence of deepfakes and video tampering detection approaches: A survey. *Multimedia Tools and Applications*, 82(7):10165–10209.

Jacob Krantz. 2024. Vln-ce: Vision-and-language navigation with continuous embeddings. https://github.com/jacobkrantz/VLN-CE. Accessed: 2024-10-20.

J. Lei, L. Zhang, M. Huang, D. Yatskar, J. Choi, Y. Zhuang, and L. Fei-Fei. 2018. Tvqa: Localized, compositional video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5486–5495.

B. Li, L. Zhu, R. Tian, S. Tan, Y. Chen, Y. Lu, Y. Cui, S. Veer, M. Ehrlich, J. Philion, and X. Weng. 2024a. Wolf: Captioning everything with a world summarization framework. *arXiv preprint*, arXiv:2407.18908.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

D. Li, Y. Liu, H. Wu, Y. Wang, Z. Shen, B. Qu, X. Niu, G. Wang, B. Chen, and J. Li. 2024c. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint*, arXiv:2410.05993.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024d. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, and L. Wang. 2024e. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. Video-llava: Learning united visual representation by alignment before projection. *Preprint*, arXiv:2311.10122.

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

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. 2024. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

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 preprint arXiv:2405.20797*.

J. Mao, X. Yang, X. Zhang, N. Goodman, and J. Wu. 2022. Clevrer-humans: Describing physical and causal events the human way. In *Advances in Neural Information Processing Systems*, volume 35, pages 7755–7768.

AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI*.

M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S.A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, and A. Oliva. 2019. Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):502–508.

Praneet Pabolu, Karan Dua, and Sriram Chaudhury. 2024a. Multi-lingual natural language generation. US Patent App. 18/318,315.

Praneet Pabolu, Karan Dua, and Sriram Chaudhury. 2024b. Multi-lingual natural language generation. US Patent App. 18/318,327.

Srikant Panda, Amit Agarwal, Gouttham Nambirajan, and Kulbhushan Pachauri. 2025. Out of distribution element detection for information extraction. US Patent App. 18/347,983.

Hitesh Laxmichand Patel, Amit Agarwal, Arion Das, Bhargava Kumar, Srikant Panda, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, and Dong-Kyu Chae. 2025. Sweeval: Do llms really swear? a safety benchmark for testing limits for enterprise use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 558–582.

Hitesh Laxmichand Patel, Amit Agarwal, Bhargava Kumar, Karan Gupta, and Priyaranjan Pattnayak. 2024. Llm for barcodes: Generating diverse synthetic data for identity documents. *arXiv preprint arXiv:2411.14962*.

V. Patraucean, L. Smaira, A. Gupta, A. Recasens, L. Markeeva, D. Banarse, S. Koppula, M. Malinowski, Y. Yang, C. Doersch, and T. Matejovicova. 2024. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36.

Priyaranjan Pattnayak, Amit Agarwal, Bhargava Kumar, Yeshil Bangera, Srikant Panda, Tejaswini Kumar, and Hitesh Laxmichand Patel. Review of reference generation methods in large language models. *Journal ID*, 9339:1263.

Priyaranjan Pattnayak, Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, and Srikant Panda. 2025a. Hybrid ai for responsive multi-turn online conversations with novel dynamic routing and feedback adaptation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 215–229.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Amit Agarwal. 2025b. Tokenization matters: Improving zero-shot ner for indic languages. *Preprint*, arXiv:2504.16977.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Bhargava Kumar, Amit Agarwal, Ishan Banerjee, Srikant Panda, and Tejaswini Kumar. 2024. Survey of large multimodal model datasets, application categories and taxonomy. *arXiv preprint arXiv:2412.17759*.

X. Shen, Y. Xiong, C. Zhao, L. Wu, J. Chen, C. Zhu, Z. Liu, F. Xiao, B. Varadarajan, F. Bordes, and Z. Liu. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024a. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. 2025. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *arXiv preprint arXiv:2504.14693*.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. 2024b. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*.

Edwin Thomas, Amit Agarwal, Sandeep Jana, and Kulbhushan Pachauri. 2025. Model augmentation framework for domain assisted continual learning in deep learning. US Patent App. 18/406,905.

Times of India. 2024. Youtube to content creators in india: We will ban such videos. Accessed: 2025-01-11.

F. Wang, X. Fu, J.Y. Huang, Z. Li, Q. Liu, X. Liu, M.D. Ma, N. Xu, W. Zhou, K. Zhang, and T.L. Yan. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Z. Wang, A. Blume, S. Li, G. Liu, J. Cho, Z. Tang, M. Bansal, and H. Ji. 2023. Paxion: Patching action knowledge in video-language foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 20729–20749.

Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. 2021. Cross-modal transferable adversarial attacks from images to videos. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15044–15053.

B. Wu, S. Yu, Z. Chen, J.B. Tenenbaum, and C. Gan. 2024a. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint*, arXiv:2405.09711.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. 2024b. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*.

B. Xie, S. Zhang, Z. Zhou, B. Li, Y. Zhang, J. Hessel, J. Yang, and Z. Liu. 2025. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer, Cham.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *CoRR*.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. 2025. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*.

Nan Yin, Mengzhu Wan, Li Shen, Hitesh Laxmichand Patel, Baopu Li, Bin Gu, and Huan Xiong. 2024. Continuous spiking graph neural networks. *arXiv preprint arXiv:2404.01897*.

H. Zhang, Y. Liu, L. Dong, Y. Huang, Z.H. Ling, Y. Wang, L. Wang, and Y. Qiao. 2023. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *arXiv preprint*, arXiv:2312.04817.

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

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *Preprint*, arXiv:2310.01852.

A Appendix

A.1 MVTamperBench Details

Dataset Name	Primary Scene Type and Unique Characteristics
STAR(Wu et al., 2024a)	Indoor actions and object interactions
PAXION(Wang et al., 2023)	Real-world scenes with nuanced actions
Moments in Time (MiT) V1(Monfort et al., 2019)	Indoor/outdoor scenes across varied contexts
FunQA(Xie et al., 2025)	Humor-focused, creative, real-world events
CLEVRER(Mao et al., 2022)	Simulated scenes for object movement and reasoning
Perception Test(Patraucean et al., 2024)	First/third-person views for object tracking
Charades-STA(for AI, 2024)	Indoor human actions and interactions
MoVQA(Zhang et al., 2023)	Diverse scenes for scene transition comprehension
VLN-CE(Krantz, 2024)	Indoor navigation from agent perspective
TVQA(Lei et al., 2018)	TV show scenes for episodic reasoning

Table 1: Summary of Datasets in MVTamperBench

Table 1 summarizes the datasets included in MVTamperBench, each contributing unique characteristics for robust tampering detection evaluation. We eliminate videos from the NTU dataset in MVTamperBench to avoid conflicting terms in the license. The dataset is then systematically expanded through the application of tampering effects to generate a comprehensive benchmark.

Table 3 summarizes MVTamperBench's unique contributions compared to existing benchmarks (Hu et al., 2024; Song et al., 2024a; Li et al., 2024e; Fu et al., 2024; Xia et al., 2024; Fei et al., 2024b) and use-cases which are spread across videos,images and documents (Agarwal et al., 2024a,b,c,d,b, 2025a; Patel et al., 2024; Agarwal and Pachauri, 2023). It also compares MVTamperBench with existing benchmarks, highlighting its focus areas, strengths, and unique contributions. We the growing interest area in MLLMS, new benchmarks that require special attention include Svbench (Yang et al., 2025), Video-MMLU (Song et al., 2025), SEA-VL (Cahyawijaya et al., 2025), MotionBench (Hong et al., 2025) and synthetic data generation techinques (Dua et al., 2024, 2025; Pabolu et al., 2024a,b).

A.2 Ablation Studies

We conduct ablation studies on key design decisions and prompt-engineering strategies to identify optimal configurations for constructing our benchmark. These experiments evaluate the effects of tampering duration, tampering position, and prompt formulation on model performance.

A.2.1 Ablation: Tampering Position and Duration

To understand the influence of tampering characteristics on model performance, we conducted two ablation studies: (1) varying the duration of tampered segments, and (2) altering their position within the video timeline.

Tampering Position Ablation. We varied the tampering position to occur after 25%, 50%, or 75% of the video timeline (Table 2). Detection performance remained largely stable across these conditions, indicating that most models are relatively insensitive to the specific location of tampering within the video. This suggests a consistent ability to maintain temporal coherence understanding regardless of when the manipulation occurs. We exclude tampering at the very beginning or end of videos, as these segments often coincide with natural scene transitions, delayed starts, or abrupt endings, which could introduce confounding noise into the benchmark.

Category	Model	25%	50%	75%
Low	Qwen2VL-7B	0.009	0.009	0.009
	LLaVaVideo-7B	0.006	0.006	0.006
	LLaVaOV-72B	0.0441	0.044	0.044
Moderate	Aria	0.7213	0.721	0.721
	Qwen2VL-72B	0.352	0.352	0.352
	Phi3.5Vision	0.707	0.707	0.707
High	VILA1.5-40B	0.879	0.879	0.879
	InternVL2.5-8B	0.721	0.7210	0.721

Table 2: Impact of tampering position on F1 scores across models with varying robustness levels. Results are reported for tampering introduced at 25%, 50%, and 75% positions within the video timeline.

Tampering Duration Ablation. We evaluated three tampering durations: 1s, 2s, and 3s. Results in Table 4 indicate a consistent improvement in detection performance with increased tampering duration across all models. Lower-performing

Benchmark	Scope (Image/Video)	Focus	Strengths	Unique Contributions
BLINK	Image	Visual reasoning	Tests spatial relationships in images.	Introduces a framework for spatial reasoning with auxiliary sketches.
MUIRBENCH	Image	Multi-image understanding	Evaluates complex real-world scenarios.	Includes multi-image relations like narrative and complementary.
MVBench	Video	Temporal reasoning, event recognition	Assesses video understanding over time.	Focuses on video dynamics with temporal task coverage.
MMBench-Video	Video	Long-form video understanding	Handles multi-step event recognition in long videos.	Evaluates LVLMs on free-form QA using temporal reasoning.
Video-MME	Video	Multi-modal video understanding	Evaluates tasks like action recognition and captioning.	Combines multiple modalities for enhanced contextual understanding.
LongVU	Video	Spatiotemporal compression	Efficiently processes long videos with adaptive compression.	Novel spatiotemporal compression mechanism using cross-modal query.
MotionEpic	Video	Object tracking	Tracks object interactions across video frames.	Implements fine-grained spatial-temporal scene graph reasoning.
Wolf	Video	Video captioning	Improves video captioning with expert strategies.	Introduces CapScore for LLM-based caption evaluation.
Sharingan	Video	Action sequence extraction	Focuses on action recognition in desktop recordings.	Proposes differential and direct frame-based methods for user action extraction.
Video-of-Thought	Video	Step-by-step video reasoning	Excels in human-like video reasoning with chain-of-thought processes.	Integrates spatial-temporal scene graphs (STSG) for fine-grained reasoning.
CARES	Video	Scene comprehension and emotional analysis	Analyzes multi-modal emotional and contextual nuances.	Integrates context-aware emotional reasoning for enhanced video understanding.
Visual-Sketchpad	Image	Visual interaction and sketching tasks	Supports creative and interactive visual reasoning.	Bridges sketch-based reasoning with image analysis for enhanced user interaction.
MovieChat	Video	Conversational video understanding	Enhances video understanding with conversational context.	Introduces dialogue-based comprehension for temporal and narrative reasoning.
MVTamperBench ***(Proposed)***	Video	Tampered video detection	Robustly identifies tampered regions in video datasets.	Unique focus on domain-specific tampering scenarios with real-world applicability.

Table 3: Enhanced comparison between Video and Image Analysis Benchmarks, with unique contributions highlighted.

Category	Model	1s	2s	3s
Low	Qwen2-VL-7B	0.009	0.201	0.257
	LLaVa-Video-7B	0.006	0.158	0.216
	LLaVa-Onevision-72B	0.044	0.257	0.249
Moderate	Aria	0.721	0.779	0.782
	Qwen2-VL-72B	0.352	0.485	0.499
	Phi3.5-Vision	0.707	0.763	0.769
High	VILA1.5-40B	0.879	0.898	0.901
	InternVL2.5-8B	0.721	0.801	0.805

Table 4: Impact of tampering duration on F1 scores across models of varying robustness. Results are reported for tampering durations of 1s, 2s, and 3s.

models (e.g., Qwen2-VL-7B, LLaVa-Video-7B) benefited more significantly, suggesting a dependency on longer anomalous intervals for effective detection. In contrast, top-performing models (e.g., VILA-1.5-40B, InternVL-2.5-8B) exhibited performance saturation, implying diminishing returns beyond a certain duration threshold.

A.2.2 Ablation: Prompt Engineering

To determine the most effective prompt formulation for evaluating MLLM robustness in video tampering detection, we conducted a series of prompt-engineering experiments. Each variant prompts the model to assess whether the video has been tampered with, as aligned with the objective of our benchmark.

- **Structured Prompt (used in benchmark)**:

Explicitly lists tampering types.
Prompt: Does this video exhibit any signs of tampering, such as corruption, blackouts, rotated frames, repeated frames, or swapped frames?

- **Generic Prompt**: Uses general phrasing that mimics natural user queries.
Prompt: Does this video exhibit any signs of tampering, manipulation, or inconsistency?

- **Chain-of-Thought (CoT) Prompt**: Instructs the model to reason step by step before deciding.
Prompt:

> *Scan the video step by step:*
> 1. *Check each segment for visual glitches, repeated or missing content, or rotations.*
> 2. *Check if any frame seems inconsistent with the rest of the video.*
> 3. *Decide if any part looks manipulated or tampered.*
>
> *Does the video show any signs of tampering, manipulation, or inconsistency?*

Category	Model	Prompt Type	F1 Score
Low	Qwen2VL-7B	Structured	0.009
		Generic	0.001
		CoT	0.001
	LLaVaVideo-7B	Structured	0.006
		Generic	0.002
		CoT	0.001
	LLaVaOV-72B	Structured	0.044
		Generic	0.001
		CoT	0.001
Moderate	Aria	Structured	0.721
		Generic	0.151
		CoT	0.148
	Qwen2VL-72B	Structured	0.352
		Generic	0.009
		CoT	0.010
	Phi3.5Vision	Structured	0.707
		Generic	0.201
		CoT	0.208
High	VILA1.5-40B	Structured	0.879
		Generic	0.458
		CoT	0.466
	InternVL2.5-8B	Structured	0.721
		Generic	0.386
		CoT	0.389

Table 5: F1 scores of models evaluated with three prompt types: Structured, Generic, and Chain-of-Thought (CoT). Structured prompts consistently yield the highest performance across all robustness categories.

Findings: Structured prompts consistently outperformed both generic and CoT prompts across all model tiers (Table 5). Low- and moderate-performing models exhibited significant performance drops with more open-ended prompts, likely due to limited temporal reasoning capabilities and lack of explicit training. Generic prompts particularly led to frequent false positives, as models misinterpreted benign variability as tampering. These findings support the use of structured prompts to ensure consistent and interpretable evaluation.

A.3 Overview of Multimodal Large Language Models (MLLMs)

Model Families and Versions Figure 13 provides a comprehensive taxonomy of Multimodal Large Language Models (MLLMs), summarizing their diversity across families, versions, and first-generation releases. The diagram branches multiple versions of a model family (e.g., *InternVL*, *LLaVA*, *VILA*, *Phi3*, *Ovis*) and separates earlier first-generation models (e.g., *Video-LLaVA* (Lin et al., 2024a; Zhu et al., 2024), *Vintern* (Doan et al., 2024), *LLama3.2-Vision* (Dubey et al., 2024; Meta, 2024)) to reflect their evolutionary development, domain-specific popularity and capabilities in conversational systems (Pattnayak et al., 2025a,b; Pattnayak et al.; Patel et al., 2025), fine-tuning (Thomas et al., 2025), and handling uses across documents (Yin et al., 2024; Agarwal et al., 2025b; Panda et al., 2025), images and videos .

This taxonomy highlights the wide range of MLLMs currently available, organized as follows:

- **InternVL Family**: Spanning models from *Intern-VL1.1* (Chen et al., 2024c; Gao et al., 2024; Chen et al., 2024b) to the advanced *Intern-VL2.5-MPO* (Chen et al., 2024d), this family emphasizes efficiency and adaptability for diverse tampering and multimodal tasks. Successive iterations demonstrate marked improvements, particularly in handling temporal disruptions such as *Dropping* and *Substitution*. The *InternVL2-5-8B* models showcases its ability to handle fine-grained spatiotemporal reasoning, highlighting its dominance in high-resource benchmarks.

- **LLaVA Family**: Starting with *LLaVa-NEXT* (Li et al., 2024d), which struggled across most benchmarks, this family has evolved with models like *LLaVa-OneVision* (Li et al., 2024b) and *LLaVa-Video* (Zhang et al., 2024),

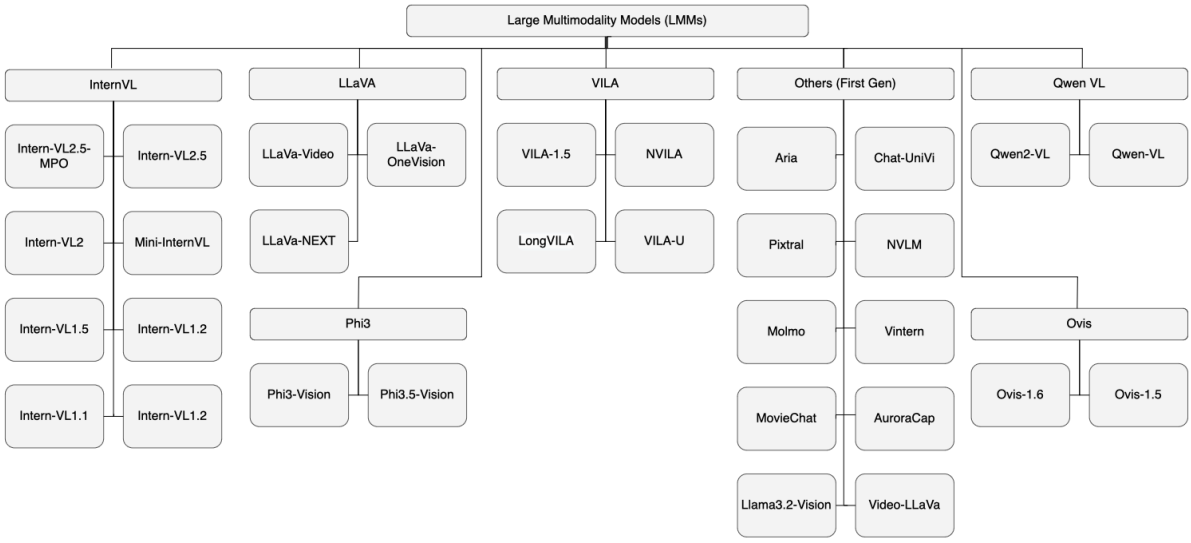


Figure 13: Taxonomy of Multimodal Large Language Models (MLLMs), organized by family, version, and first-generation releases.

demonstrating significant improvements in task-specific video understanding through optimized pretraining and alignment techniques. Despite the advancements, *LLaVa-OneVision & LLaVa-Video* continues to face challenges in handling complex temporal disruptions, unlike *Chat-UniVi*, which has emerged as a robust alternative.

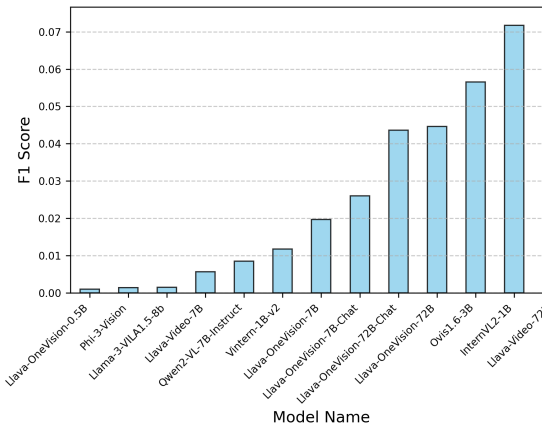
- **VILA Family**: The *VILA* series, including *VILA-1.5* (Lin et al., 2024b), *LongVILA* (Xue et al., 2024), and *VILA-U* (Wu et al., 2024b), demonstrates exceptional overall performance due to robust training pipelines and innovative architectures. Notably, *VILA-40B* excels across benchmarks, which demonstrates advanced fine-grained and long-form video understanding capabilities, particularly in addressing tampering types like *Masking* and *Rotation*. Its architectural design allows it to efficiently process long-context visual inputs, setting a new standard for performance in large-scale benchmarks.

- **Phi3 Family**: Known for its scalable architecture and focus on real-world applicability, the *Phi3-Vision* (Abdin et al., 2024) models struggle across tampering scenarios. The *Phi3.5-Vision* model, in particular, highlights the benefits of improved tokenization and training strategies. It outperforms it's earlier versions by leveraging better alignment between visual and textual modalities.

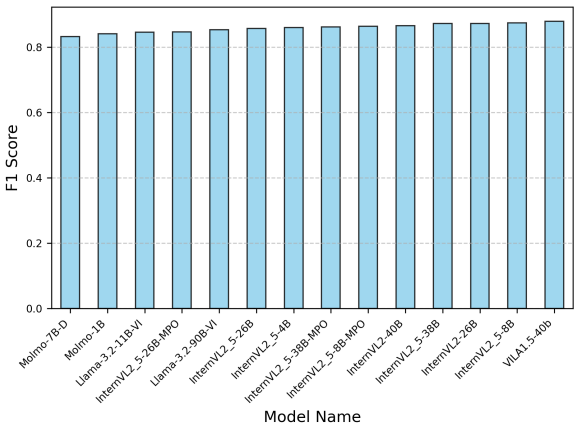
- **Ovis Family**: Optimized for fine-grained visual-text alignment, *Ovis* (Lu et al., 2024) models leverage specialized datasets for tasks requiring high-resolution image interpretation and contextual reasoning. While the smaller version struggled with temporal coherence, subsequent larger model versions show promising improvements in spatial reasoning tasks like *Masking* and *Rotate*. Such enhancement results from better alignment between text and vision modalities.

- **First-Generation Releases**: Models such as *Chat-UniVi* (Jin et al., 2024), *Molmo* (Deitke et al., 2024), *NVLM* (Dai et al., 2024), and *Pixtral* (Agrawal et al., 2024) represent earlier efforts in video-language modeling. While some, like *Molmo* and *Aria* (Li et al., 2024c), continue to show competitive performance due to innovative training strategies, others, such as *NVLM*, are limited by suboptimal optimization for temporal reasoning, which hinders their ability to adapt to tampering scenarios.

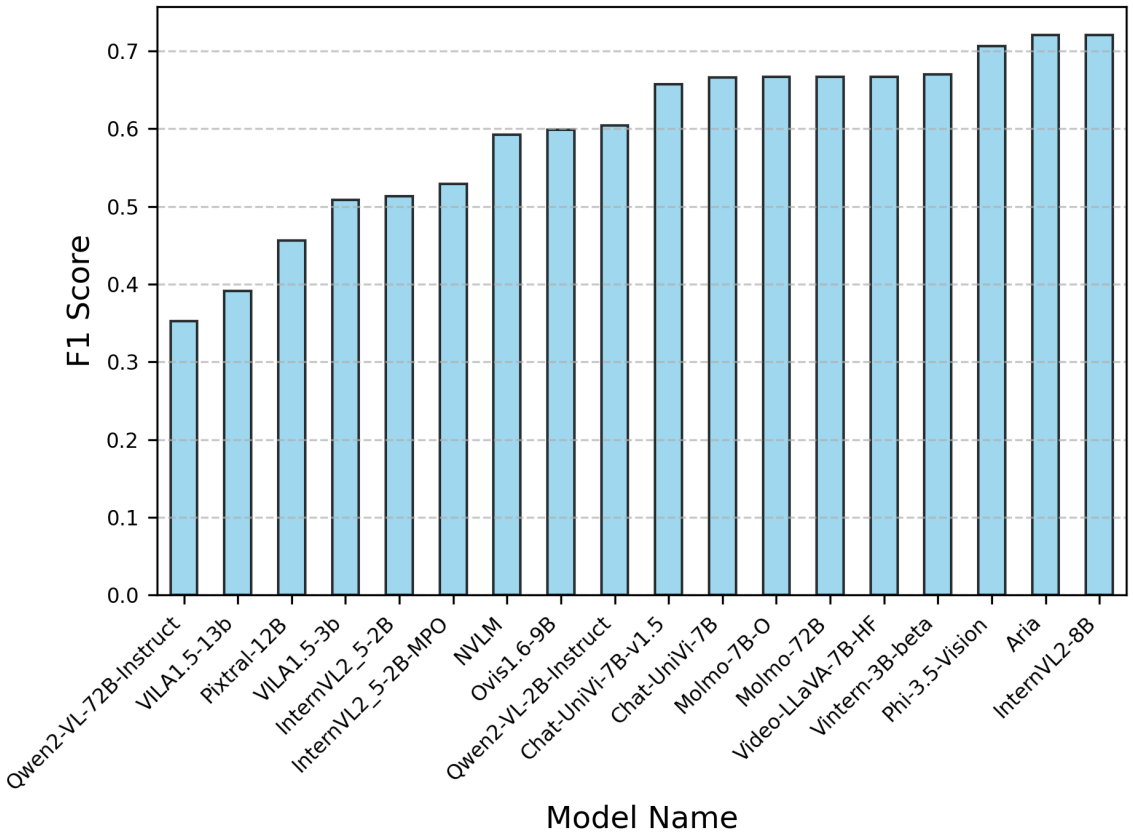
- **Qwen-VL Family**: The *Qwen2-VL* (Wang et al., 2024b) and *Qwen-VL* (Yang et al., 2024) models are recent entrants that combine advanced architectures with scalable parameterization. They achieve strong results in grounding and visual reasoning tasks but struggle in detecting tampering across task and scenarios. Scaling the model size does help the performance but is still below the average performance of models in the study.



(a) MLLMs in the Low Performing Category



(b) MLLMs in the High Performing Category



(c) MLLMs in the Moderate Performing Category

Figure 14: F1 (overall) performance of individual models across low, moderate, and high-performing categories. Models like *InternVL-2.5* lead high-performing groups, while *Llava-OneVision* models underperform consistently.

A.4 Extended Results & Analysis

A.4.1 Analysis based on Performance Categories

Figure 14 examines individual model performance across all categories, providing insights into trends among model families, versions, and architectures.

For low-performing models (Figure 14a), **Llava-OneVision** exhibits consistently weak performance across tampering types, even at larger parameter sizes (e.g., Llava-OneVision-72B). This suggests potential architectural and training data limitations, particularly for temporal coherence tasks. Interestingly, **Qwen2-VL-7B** underperforms significantly compared to its larger counterpart, **Qwen2-VL-72B**, which achieves a notable improvement. This indicates that increasing model

size, combined with its training paradigm, positively impacts robustness for this family.

In the moderate-performing category (Figure 14c), several trends emerge. **Llama3.2-11B** and **Llama3.2-90B** exhibit very similar performance despite the significant increase in size. These models, trained on the same recipe and dataset, highlight that merely increasing parameter count does not drastically enhance tampering detection capabilities. Another interesting observation is the improvement shown by **Phi3.5-Vision** and **Vintern-Beta** over their predecessors (**Phi3-Vision** and **Vintern**). This suggests that targeted architectural modifications or additional task-specific training significantly contribute to robustness.

For high-performing models (Figure 14b), **InternVL-2.5** dominates across tampering effects, with smaller versions (e.g., 4B) achieving comparable performance to their larger counterparts. This demonstrates that efficient architectural design and training strategies can offset limitations in model size. **VILA1.5-40B**, a model specifically designed for long-form video understanding, showcases exceptional robustness, emphasizing the importance of task-specific optimization. Notably, **Molmo-72B** and **NVLM-72B** exhibit below-average performance relative to other large models or their smaller counter parts, indicating inefficiencies in parameter utilization or potential overfitting to pretraining data and tasks.

Another noteworthy observation across categories is the consistency within certain model families. For example, the **Llava-Video** and **Chat-UniVi** models outperform their Llava-OneVision counterparts, demonstrating the importance of video-specific training for tampering detection. Conversely, while **Molmo-1B** excels among small models, its larger variant (**Molmo-72B**) does not scale proportionally in performance, reinforcing the need for efficient scaling and training strategies.

A.4.2 Analysis based on Model Architectures

To better understand model performance under video tampering, we conducted a comparative analysis of MLLMs' architectural design, alignment strategies, and training paradigms. Table 6 summarizes key configurations across models stratified by their performance category (low, moderate, high).

Our findings suggest that robust MLLMs typically adopt deeper integration strategies between vision and language modalities. High-performing models like **VILA**, **InternVL**, and **Aria** utilize multi-stage training pipelines, explicit visual-text alignment layers (e.g., projectors or middleware), and instruction tuning on curated or human-annotated datasets. In contrast, moderate-performing models often rely on lightweight fusion (e.g., MLPs or prompt tuning) or lack post-alignment tuning stages. Low-performing models tend to use frozen CLIP encoders and shallow decoders with minimal multimodal alignment.

These insights support the hypothesis that robustness to spatiotemporal manipulations correlates with stronger cross-modal alignment and iterative supervision across training stages.

A.4.3 Analysis based on Model Size

A closer examination of individual model trends across size categories (Figure 15) reveals several noteworthy patterns.

Small Models (<7B). Among small models (Figure 15a), **Molmo-1B** demonstrates exceptional robustness, outperforming several medium-sized models and achieving consistency across all tampering types. Another notable small model, **Phi3.5-Vision**, shows drastic improvement over its predecessor, **Phi3-Vision**, highlighting the impact of architectural updates and extended task-specific training.

For the **VILA model family**, **VILA1.5-3B** performs better than many other small models, showcasing the benefits of targeted optimization for long-form video understanding. However, its performance lags behind **Molmo-1B** and **Phi3.5-Vision**, indicating room for improvement in handling complex tampering scenarios.

Interestingly, **InternVL-2.5-4B** matches or exceeds the performance of several medium-sized models, emphasizing the importance of efficient design over raw parameter counts. Models like **Llava-OneVision-7B**, however, remain among the weakest performers, suggesting that training approaches and architectural focus on static data limit their ability to handle tampering.

Medium Models (7B–26B). For medium-sized models (Figure 15c), we observe considerable variability. Variants such as **Molmo-D** and **Molmo-O** exhibit similar performance, suggesting limited scalability within the family. Conversely, models like **InternVL-2.5-8B** maintain exceptional robustness, outperforming even larger models in the same category.

Perf.	Model	Vision	Language	Alignment Strategy	Training Strategy
High	Aria	SigLIP-400M	Aria-MoE	MoE decoder + lightweight ViT	4-stage: language, multimodal, long-context, post-training
High	LLaMA	In-House	LLaMA-3.2	—	Iterative SFT, rejection sampling, DPO on curated data
High	VILA	InternViT-6B	Yi-34B	Deep embedding projection + joint tuning	3-stage: projector init, visual pretrain, instruction tuning
High	InternVL	InternViT-6B v2.5	Qwen2.5	QLLaMA middleware harmonization	Contrastive → generative → supervised tuning
Moderate	Vintern	InternViT-300M	Qwen2.5	MLP projector w/ visual instruction tuning	2-stage: full-param + LoRA, cross-entropy loss
Moderate	Qwen2-VL	Custom ViT	Qwen2	Dynamic resolution + M-RoPE	3-stage: image–text pretrain + instruction tuning
Moderate	Pixtral	ViT-400M	Nemo-12B	Decoder fusion, ROPE-2D, sequence packing	Interleaved image–text pretraining
Moderate	NVLM	InternViT-6B	Qwen2	Decoder/cross-attn/hybrid variants	Pre-align (frozen LLM), then multimodal SFT
Moderate	Molmo	CLIP ViT-L/14	Qwen2-72B	Multi-crop connector + decoder LLM	Length-conditioned captions → multitask tuning
Moderate	Chat-UniVi	CLIP-ViT-L/14-336	Vicuna-v1.5-7B	Unified visual tokens + multi-scale semantics	2-stage: frozen pretrain + joint instruction tuning
Low	PHI	CLIP ViT-L/14	Phi-3	Transformer decoder (block-sparse)	2-phase: filtered web + synthetic

Table 6: Architecture, alignment, and training strategies for MLLMs stratified by robustness category.

Pixtral, a medium-sized model that claims higher performance on other benchmarks, delivers average results here, highlighting that tampering-specific robustness requires distinct optimization strategies. Similarly, **Chat-UniVi** demonstrates an advantage over other Llava-family models, reinforcing the role of video-specific training in achieving higher tampering detection performance.

In the **VILA model family**, **VILA1.5-8B** underperforms compared to both smaller and larger VILA models, making it an interesting anomaly. This drop in performance could stem from suboptimal parameter tuning or an architectural bottleneck that affects scalability. Meanwhile, **VILA1.5-13B** and **VILA1.5-40B** continue to improve with increasing size, emphasizing the scalability of this family for long-form video tampering tasks.

Large Models (>26B). Among large models (Figure 15b), we observe mixed performance trends. While **InternVL-2.5** continues to dominate, its smaller variants (e.g., 4B, 8B) achieve comparable results, raising questions about the marginal benefits of scaling up within this family. **VILA1.5-40B**, specifically designed for long-form video understanding, ranks among the best-performing large models, showcasing the value of video-specific training & optimization.

The **Qwen2-VL** family highlights the importance of scaling when paired with architectural optimization. **Qwen2-VL-72B** significantly outperforms its 7B counterpart, demonstrating the advantages of increased parameter counts and more extensive pretraining. However, its performance still lags behind other large models like **InternVL-2.5-40B** and **VILA1.5-40B**, suggesting potential inefficiencies in architecture or video-specific optimization.

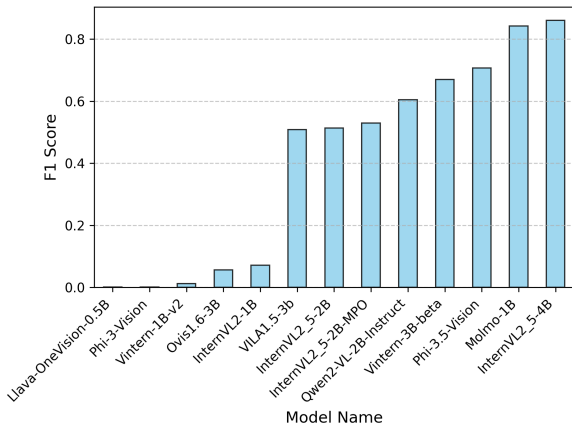
The **Llama3.2-Vision** family offers a nuanced perspective on scaling. While **Llama3.2-11B** and **Llama3.2-90B** exhibit very similar performance, underscoring that increasing parameter count alone does not yield proportional performance gains. This trend reflects the importance of complementing scaling with architectural innovations and diverse, tamper-focused training data.

While **InternVL-2.5** continues to dominate, its smaller variants (e.g., 4B, 8B) achieve comparable results, raising questions about the marginal benefits of scaling up within this family. Similarly, **Molmo-72B**, despite its size, fails to match the performance of its smaller counterparts, indicating inefficiencies in parameter utilization or overfitting to pretraining data.

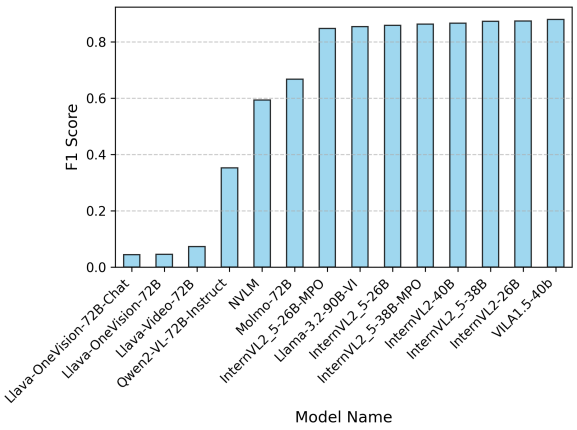
In contrast, **Qwen2-VL-72B** delivers subpar results relative to its size but still outperforms its smaller sibling, **Qwen2-VL-7B**. This highlights that while scaling can improve performance, architectural and training advancements are critical for leveraging the full potential of larger models.

A.4.4 Analysis across Video Task Types

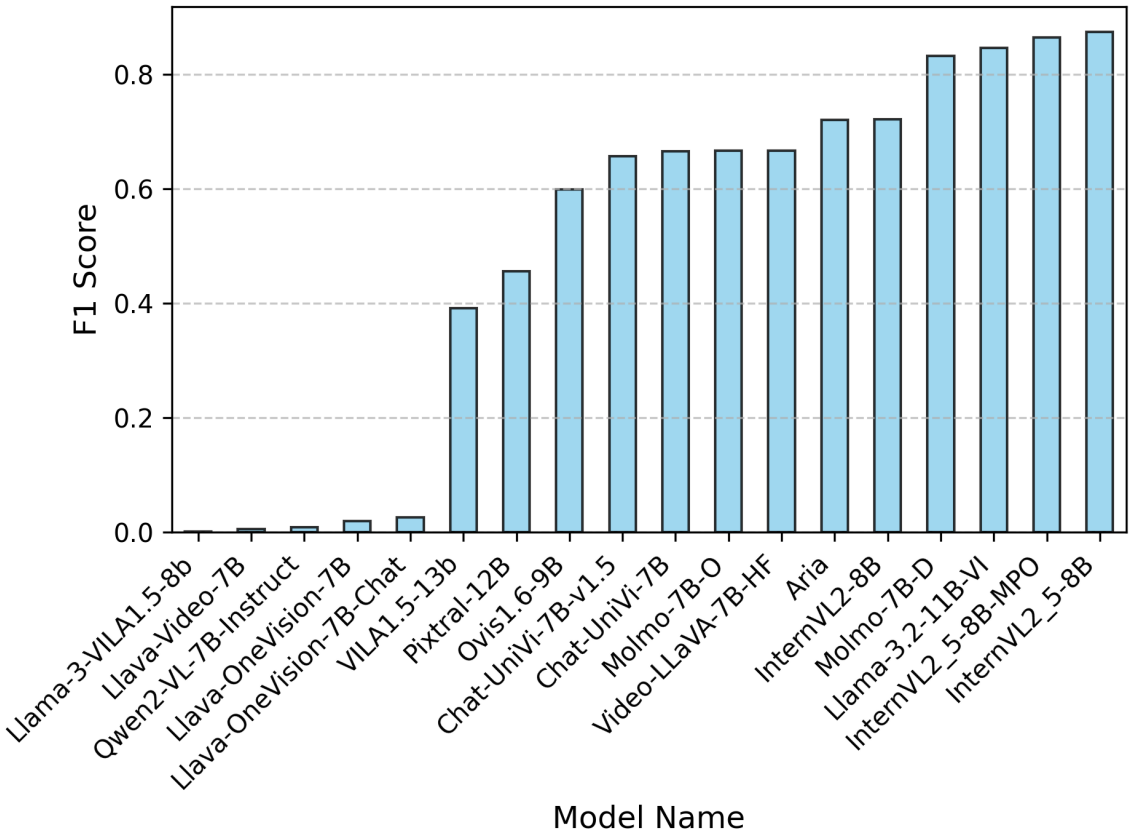
Insights across Tampering Types. Figure 16 reveals nuanced differences in performance across tampering types and video tasks. *Masking* consistently emerges as the least disruptive effect, particularly in tasks like **Moving Direction** and **Ob-**



(a) Performance Distribution of Small(<7B) MLLMs



(b) Performance Distribution of Large (>26B) MLLMs



(c) Performance Distribution of Medium(>=7B & <=26B) MLLMs

Figure 15: F1 (overall) performance of individual models across small, medium, and large model size categories. Models like *InternVL-2.5* lead high-performing groups, while *Llava-OneVision* models underperform across categories.

ject Existence, where models rely more on contextual cues than on fine-grained spatial details. Conversely, *Dropping* and *Repetition* create the most significant challenges, particularly in tasks involving long-term temporal dependencies, such as **Action Sequence** and **Counterfactual Inference**.

Notably, the performance trends also vary across model categories. High-performing models demonstrate consistent robustness across all tasks and tampering types, while low-performing models exhibit the highest susceptibility to temporal and spatial disruptions. Moderate-performing models, on the other hand, display a mix of strengths and weaknesses, excelling in tasks with static or localized

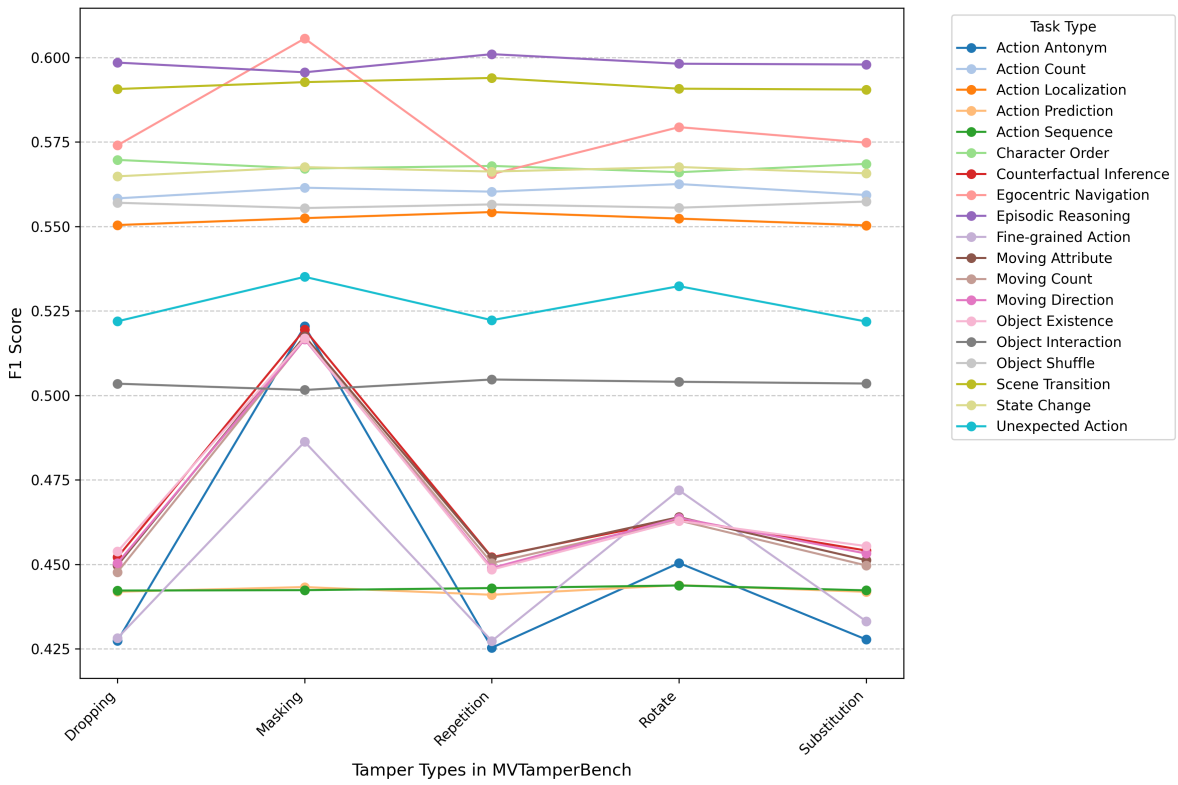


Figure 16: F1 (overall) scores across tampering types for task categories. *Masking* is less disruptive, while *Dropping* and *Substitution* degrade performance in complex tasks like *Counterfactual Inference*.

changes but struggling with tasks requiring temporal coherence.

Trends across Task Categories. Examining the task-wise trends, several key observations emerge:

- Tasks involving **long-term temporal reasoning** (e.g., **Counterfactual Inference, Action Sequence**) are more sensitive to tampering, particularly *Dropping*, *Repetition*, and *Substitution*, which disrupt narrative continuity.
- Tasks with **localized changes** (e.g., **State Change, Episodic Reasoning**) are less affected by tampering, as models rely on static visual cues.
- Tasks requiring **fine-grained spatial understanding** (e.g., **Fine-Grained Action**) show significant degradation under *Rotation*, indicating a limitation in models' ability to process spatial distortions.
- Performance disparities across tampering types highlight architectural strengths and weaknesses. For instance, **InternVL-2.5** excels in tasks like **Action Prediction** and **Counterfactual Inference**, owing to its advanced temporal reasoning capabilities.

A.5 Benchmarking Efforts

Our benchmarking efforts encompass 45 models across diverse categories and tampering scenarios, including Drop, Mask, Repeat, Rotate, and Substitute (Table 7). The results reveal significant variability in performance, influenced by model size, architecture, and training data.

Models such as *VILA1.5-40B* and *InternVL2.5-8B* emerged as top performers, achieving consistent resilience across all tampering types, with overall scores of 0.879 and 0.875, respectively. This highlights the importance of architectural innovations and advanced training techniques for tampering robustness. In contrast, early-generation models like *LLaVA-OneVision* underperformed across all categories, with overall scores as low as 0.001, reflecting limitations in temporal coherence and token alignment.

Specialized models for video tasks, such as *Chat-UniVi* and *Video-LLaVA*, demonstrated substantial improvements over base *LLaVA* models. *Chat-UniVi-7B-v1.5* achieved an overall score of 0.658, significantly outperforming *LLaVA-OneVision*, showcasing its ability to handle complex temporal manipulations. Meanwhile, *Video-LLaVA-7B-HF* maintained robust performance

across categories, further validating the effectiveness of unified tokenization and video-specific optimizations. However, *LLaVA-Video*, despite efforts to improve alignment and pretraining, continues to struggle with certain tampering types, reflecting the challenges of adapting image-centric architectures to video modalities.

Interestingly, medium-sized models like *Phi3.5-Vision* demonstrated notable performance improvements compared to earlier iterations such as *Phi3-Vision*, indicating that scaling alone does not account for robustness gains. Specialized models like *Ovis1.6-Gemma2-9B* showcased strengths in spatial tampering scenarios (e.g., Mask and Rotate) but struggled with temporal disruptions like Repeat and Substitute. This trend underscores the importance of task-specific optimizations.

Future Benchmarking Plans. While our analysis has covered an extensive set of models, several promising entries are yet to be evaluated. Models such as *NVILA* (Liu et al., 2024), *LongVILA* (Xue et al., 2024), and *AuroraCap* (Chai et al., 2024) are currently being integrated into our benchmarking framework, with active collaborations underway to ensure seamless evaluations. Early insights suggest that these models could offer competitive performance in handling long-form video tampering and multimodal reasoning.

Additionally, we plan to expand the scope of evaluations for models like *InternVL-1*, *MovieChat* (Song et al., 2024a,b), *Vintern*, and future iterations of *Chat-UniVi* and *Video-LLaVA*. While *Chat-UniVi* has shown impressive robustness across temporal and spatial tampering scenarios, and *Video-LLaVA* continues to improve, exploring these models under additional tampering techniques will provide deeper insights into their limitations and areas for refinement.

The dynamic and evolving nature of our benchmarking framework ensures that future evaluations will continue to capture advancements in architectural design and tampering robustness.

A.6 Key Findings

Consistent Performers. The *InternVL-2.5* series consistently outperforms other models by achieving strong F1 (overall) scores across all tampering types. Notably, even smaller variants like *InternVL-2.5-4B* match the robustness of larger models. Such results highlight the efficiency of its architecture and training strategy.

Effect-specific Strengths. Models such as *Phi3.5-Vision* excel in detecting *Masking* tampering, which indicates its specialized capabilities for handling visual obfuscations. Similarly, the *VILA1.5-40B*, designed for long-form video understanding, excels in spatial-temporal tasks.

Weaker Models. Certain models, including *Llava-OneVision* variants, exhibit consistent weaknesses, particularly with temporal disruptions like *Dropping* and *Repetition*. This result may suggest limitations in their architectural designs and training paradigms.

Tampering Insights. *Dropping* and *Repetition* emerge as the most challenging tampering types for all model categories, reflecting the difficulty of maintaining temporal coherence under such manipulations. In contrast, *Masking* is relatively less disruptive, particularly for tasks relying on contextual cues.

A.7 Discussion and Future Directions

The findings shed light on the importance of task-specific optimization and tamper-aware training for improving model robustness. Tasks that involve complex temporal dependencies or fine-grained spatial reasoning highlight critical gaps in current architectures, which need to better integrate temporal embeddings and multi-scale spatial attention mechanisms. MVTamperBench highlights the critical importance of robust architectures and diverse training data & strategies for achieving tampering resilience. Below, we outline actionable insights and avenues for future exploration:

Expanding Benchmark Scope. To enhance the benchmark's comprehensiveness, future iterations could evaluate additional model families, including emerging MLLMs. This will ensure a broader understanding of robustness trends.

Addressing Weak Models. Models like *Llava-OneVision* & *Qwen-2-VL* highlight the need for targeted improvements. Techniques such as adversarial training, task-specific fine-tuning, and architectural enhancements could improve performance.

Introducing New Tampering Types. Expanding the benchmark to include tampering techniques such as localized masking, noise injection, and frame-level shuffling would provide a more nuanced evaluation of model resilience. Additionally,

exploring domain-specific tampering types for critical applications like healthcare and surveillance could reveal context-dependent vulnerabilities.

Task-specific Insights. Examining performance at a finer granularity—e.g., evaluating models on specific task categories or within specialized domains—could provide actionable guidance for training and optimization.

Scaling Considerations. Observations such as diminishing returns for models like *Molmo-72B* and *Llama-3.2-90B-Vision* emphasize the need for efficient scaling strategies. Future work could explore methods for optimizing parameter utilization and balancing architectural complexity with training data diversity.

Integration with Real-world Applications. Extending MVTamperBench to evaluate tampering resilience in real-world domains like media verification, misinformation detection, and legal forensics could uncover application-specific challenges and provide a pathway for practical deployments.

By exploring these directions, we hope that our MVTamperBench will evolve as a cornerstone benchmark for tampering detection, and further drive innovation in tamper-resilient MLLMs and foster trust in their real-world applications.

Model	Size	Drop	Mask	Repeat	Rotate	Substitute	Overall	Performance Category	Size Category
Phi-3-Vision	4	0.002	0.002	0.000	0.002	0.002	0.001	Low	Small
llava-onevision-qwen2-0.5b-ov	1	0.001	0.001	0.001	0.001	0.001	0.001	Low	Small
Llama-3-VILA1.5-8b	8	0.001	0.003	0.001	0.003	0.001	0.002	Low	Medium
llava_video_qwen2_7b	7	0.006	0.006	0.005	0.005	0.006	0.006	Low	Medium
Qwen2-VL-7B-Instruct	7	0.005	0.004	0.004	0.025	0.005	0.009	Low	Medium
Vintern-1B-v2	1	0.013	0.011	0.012	0.011	0.011	0.012	Low	Small
llava-onevision-qwen2-7b-ov	7	0.016	0.038	0.010	0.019	0.016	0.020	Low	Medium
llava-onevision-qwen2-7b-ov-chat	7	0.020	0.050	0.013	0.027	0.020	0.026	Low	Medium
llava-onevision-qwen2-72b-ov-chat	72	0.021	0.125	0.018	0.034	0.021	0.044	Low	Large
llava-onevision-qwen2-72b-ov	72	0.020	0.132	0.017	0.033	0.021	0.045	Low	Large
Ovis1.6-Llama3.2-3B	3	0.057	0.064	0.050	0.053	0.058	0.057	Low	Small
InternVL2-1B	1	0.074	0.070	0.067	0.074	0.074	0.072	Low	Small
llava_video_qwen2_72b	72	0.019	0.254	0.017	0.053	0.022	0.073	Low	Large
Qwen2-VL-72B-Instruct	72	0.279	0.544	0.282	0.374	0.284	0.352	Moderate	Large
VILA1.5-13b	13	0.378	0.422	0.383	0.391	0.382	0.391	Moderate	Medium
Pixtral-12B	12	0.452	0.469	0.453	0.454	0.453	0.456	Moderate	Medium
VILA1.5-3b	3	0.457	0.609	0.469	0.536	0.470	0.508	Moderate	Small
InternVL2_5-2B	2	0.500	0.538	0.505	0.523	0.499	0.513	Moderate	Small
InternVL2_5-2B-MPO	2	0.518	0.558	0.519	0.534	0.517	0.529	Moderate	Small
NVLM	72	0.588	0.595	0.598	0.595	0.588	0.593	Moderate	Large
Ovis1.6-Gemma2-9B	9	0.600	0.593	0.600	0.602	0.600	0.599	Moderate	Medium
Qwen2-VL-2B-Instruct	2	0.610	0.594	0.606	0.601	0.612	0.605	Moderate	Small
Chat-UniVi-7B-v1.5	7	0.662	0.642	0.663	0.661	0.661	0.658	Moderate	Medium
Chat-UniVi-7B	7	0.667	0.666	0.666	0.666	0.666	0.666	Moderate	Medium
molmo-7B-O-0924	7	0.667	0.667	0.667	0.667	0.667	0.667	Moderate	Medium
Video-LLaVA-7B-HF	7	0.667	0.667	0.667	0.667	0.667	0.667	Moderate	Medium
molmo-72B-0924	72	0.667	0.667	0.667	0.667	0.667	0.667	Moderate	Large
Vintern-3B-beta	3	0.670	0.669	0.674	0.669	0.670	0.670	Moderate	Small
Phi-3.5-Vision	4	0.676	0.822	0.677	0.682	0.677	0.707	Moderate	Small
InternVL2-8B	8	0.705	0.761	0.689	0.740	0.711	0.721	Moderate	Medium
Aria	25	0.717	0.738	0.716	0.719	0.716	0.721	Moderate	Medium
molmo-7B-D-0924	7	0.833	0.833	0.832	0.833	0.833	0.833	High	Medium
molmoE-1B-0924	1	0.842	0.842	0.842	0.842	0.842	0.842	High	Small
Llama-3.2-11B-Vision-Instruct	11	0.846	0.847	0.848	0.847	0.845	0.847	High	Medium
InternVL2_5-26B-MPO	26	0.848	0.848	0.847	0.848	0.848	0.848	High	Large
Llama-3.2-90B-Vision-Instruct	90	0.854	0.853	0.853	0.853	0.854	0.853	High	Large
InternVL2_5-26B	26	0.858	0.858	0.858	0.858	0.858	0.858	High	Large
InternVL2_5-4B	4	0.860	0.861	0.860	0.861	0.860	0.860	High	Small
InternVL2_5-38B-MPO	38	0.857	0.868	0.859	0.868	0.859	0.862	High	Large
InternVL2_5-8B-MPO	8	0.864	0.864	0.864	0.864	0.864	0.864	High	Medium
InternVL2-40B	40	0.863	0.870	0.864	0.870	0.865	0.866	High	Large
InternVL2_5-38B	38	0.871	0.875	0.872	0.875	0.871	0.873	High	Large
InternVL2-26B	26	0.872	0.876	0.872	0.874	0.873	0.873	High	Large
InternVL2_5-8B	8	0.875	0.875	0.875	0.875	0.875	0.875	High	Medium
VILA1.5-40b	40	0.879	0.880	0.878	0.880	0.879	0.879	High	Large

Table 7: Performance metrics for various models.