# VidNum-1.4K: A Comprehensive Benchmark for Video-based Numerical Reasoning

Shaoyang Cui*
sy-cui@thu.edu.cn
Department of Psychological and Cognitive Sciences,
Tsinghua University
Beijing, Beijing, China

Lingbei Meng*
250010166@slai.edu.cn
Shenzhen Loop Area Institute
Shenzhen, Guangdong, China

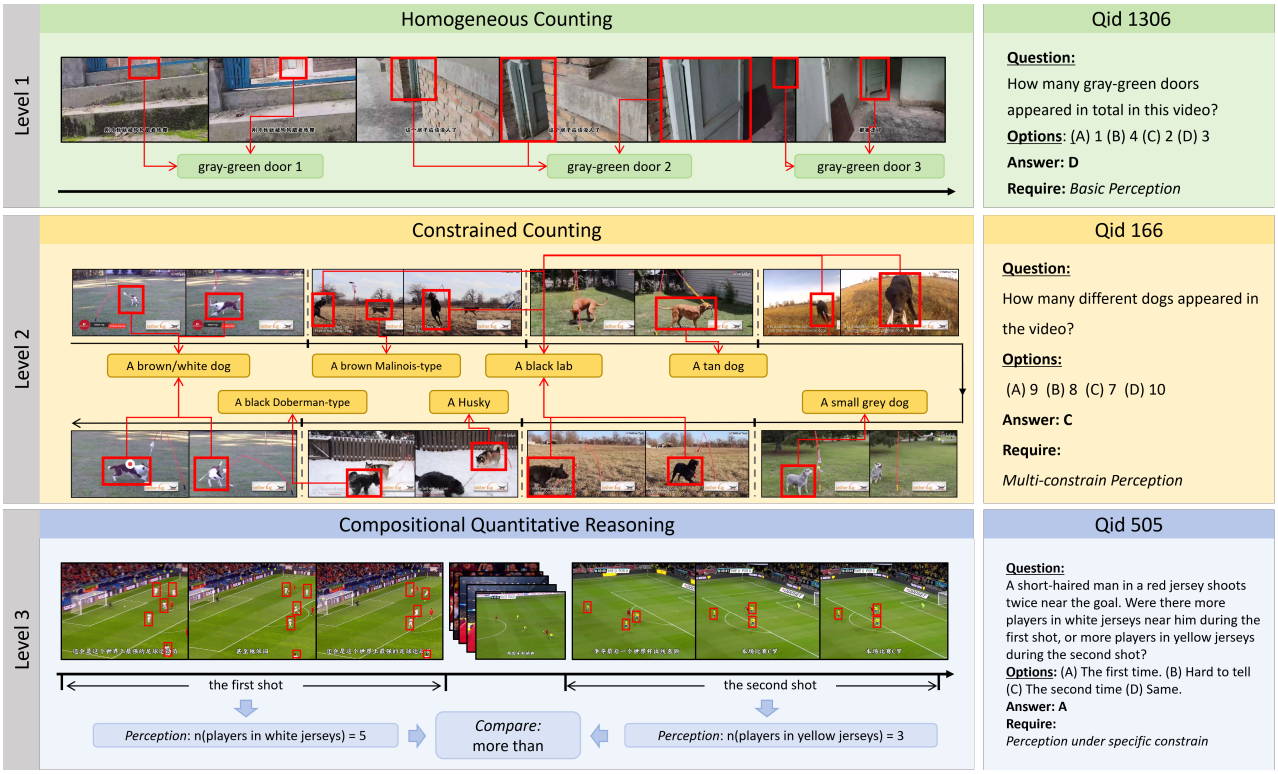


**Figure 1: Demo of Three-Level Questions in VidNum1.4K.**

## Abstract

Video-based numerical reasoning provides a premier arena for testing whether Vision-Language Models (VLMs) truly "understand" real-world dynamics, as accurate numerical deduction necessitates a profound grasp of temporal events, object permanence, and compositional logic beyond superficial pattern matching. However, existing benchmarks are often confined to narrow domains, such as repetitive athletic motions, or treat simple counting merely as a superficial regression task, failing to assess multi-step numerical logic within the inherent complexity of real-world multimedia content. We introduce ***VidNum-1.4K***, a comprehensive VideoQA benchmark comprising 1,379 strictly human-annotated video-question pairs designed to evaluate genuine numerical reasoning across highly diverse environments, encompassing object, action, and event quantification. The ***VidNum-1.4K*** is uniquely structured into a three-level hierarchy that evolves from direct visual perception to video-based compositional numerical reasoning, requiring models to perform arithmetic operations, comparisons, and logical deductions grounded in temporal evidence. Our evaluations across a diverse suite of state-of-the-art VLMs reveal a striking reasoning gap: while the Gemini-3.1-pro barely reach a 60% accuracy threshold, representative open-source families struggle heavily in the 25%–45% range. These findings demonstrate that current VLMs still lack a stable "internal world model", positioning *VidNum-1.4K*as a demanding diagnostic testbed for the next generation of numerical video intelligence.

## CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Question answering**; *Computer vision*; *Knowledge representation and reasoning*.

*Both authors contributed equally to this research.

## Keywords

Visual Language Model, Video Question-Answering, Video Reasoning

## 1 Introduction

Vision-language models (VLMs) have achieved remarkable progress across a wide range of video understanding tasks, evolving rapidly from contrastive pretraining to instruction-following multimodal assistants (e.g., CLIP, Flamingo, and LLaVA-style systems) [2, 17, 19]. However, as these models grow increasingly capable, a critical scientific question remains open: do current VLMs genuinely build a stable **World Model** to understand dynamic scenes, or do they merely memorize dataset priors and superficial correlations from their pretraining?

We argue that **video-based numerical reasoning** serves as an excellent arena to answer this question. By video-based numerical reasoning, we refer to tasks that include counting and basic calculations or logical deductions regarding numbers and object quantities in videos. Unlike standard open-ended semantic QA, this task not only tests a model's foundational object recognition and event segmentation capabilities within continuous streaming inputs, but it strictly demands that the model truly *understands* the scenes, events, and underlying concepts. To successfully track *what* appears, *when* it happens, and *how* quantities evolve, a model must possess a genuine understanding of the physical *world*, as shortcut language priors are highly ineffective in this setting.

Existing benchmarks partially cover this challenge but remain fragmented. In standard video QA benchmarks such as TGIF-QA [12], NExT-QA [24], and ActivityNet-QA [26], counting is included but only as a minor subset within broader semantic QA objectives, lacking an explicit hierarchy of quantitative difficulty. Conversely, dedicated repetition-counting benchmarks like QUVA Repetition [20], Countix/RepNet [7], and UCFRep [29] primarily evaluate cycle estimation for repeated actions as a regression-style task. While valuable, these resources do not jointly assess multi-type counting (object/event/action) and count-conditioned inference across videos with frequent shot cuts. The core limitation in the field is the lack of a complete evaluation protocol for compositional numerical reasoning.

To address this gap, we present *VidNum-1.4K*, a dedicated multiple-choice benchmark for numerical reasoning in video QA. The benchmark comprises 1,379 multiple-choice questions (MCQs), each based on an independent video clip. To ensure comprehensive evaluation, the videos are drawn from highly diverse sources—including real-world, documentary/educational, and virtual-world footage—and vary significantly in length.

In summary, our main contributions are:

- We introduce *VidNum-1.4K*, a dedicated and diverse video QA benchmark designed specifically to evaluate the numerical reasoning and "world understanding" capabilities of VLMs.
- We conduct a comprehensive empirical evaluation of state-of-the-art open-source and closed-source VLMs on VidNum-1.4K. Our exhaustive tests reveal a significant performance gap, indicating that robust numerical understanding in videos remains far from solved for current models.

## 2 Related Works

### 2.1 Video Understanding Tasks

Video understanding encompasses diverse tasks like action recognition, retrieval, and QA, driven by benchmarks targeting motion, temporal grounding, and semantics (e.g., Kinetics [14], ActivityNet [11], Something-Something V2 [10], and MSR-VTT [25]). While datasets such as TGIF-QA [12] and NExT-QA [24] advance temporal and causal reasoning, they lack a focus on dedicated numerical competence. Meanwhile, static image benchmarks like CLEVR [13] and TallyQA [1] demonstrate the inherent difficulty of numerical reasoning without temporal dynamics. Together, these gaps highlight the necessity of a specialized numerical video benchmark.

### 2.2 Existing Video Counting Benchmarks

Current video counting resources largely fall into two categories. The first focuses on cycle regression for repetitive actions (e.g., QUVA Repetition, Countix/RepNet, and UCFRep [7, 20, 29]), prioritizing temporal periodicity modeling over language-grounded, multi-step reasoning. The second embeds counting within general VideoQA (e.g., TGIF-QA, NExT-QA, and ActivityNet-QA [12, 24, 26]), where quantitative control is secondary. This limits their utility for diagnosing count-conditioned arithmetic or cross-event numerical consistency. Additionally, while event-centric datasets like SoccerNet and UCF-Crime [9, 22] contain frequency statistics, they are designed for action spotting and anomaly detection, rather than compositional numerical reasoning.

### 2.3 Vision-Language Models

While early video understanding relied on specialized visual backbones for classification and spatio-temporal modeling (e.g., Two-stream [21], C3D [23], I3D [6], SlowFast [8], and TimeSformer [5]), the paradigm has recently shifted toward instruction-following Vision-Language Models (VLMs). Foundational models like CLIP [19] established scalable alignment, paving the way for robust in-context reasoning (Flamingo [2], BLIP-2 [15]) and interactive multimodal systems (LLaVA [17], LLaVA-NeXT [16]). For video, recent VLMs extend these capabilities to temporal understanding (Video-LLaMA [28], Video-ChatGPT [18], VideoLLaMA3 [27]). Furthermore, frontier general-purpose models, such as the Qwen-VL family [3, 4], demonstrate strong broad-domain performance. Despite this progress, existing evaluations still lack a fine-grained protocol for numerical video reasoning—the exact gap VidNum1.4K addresses.

## 3 The VidNum-1.4K

As illustrated in Figure 2, VidNum-1.4K comprises 1,379 meticulously curated multiple-choice questions (MCQs), spanning three primary counting targets: Object (647), Action (343), and Event (389). Each question is grounded in a distinct video clip ranging from 5 to 120 seconds in duration. To facilitate a fine-grained diagnostic evaluation of VLMs, we structure the benchmark into a three-level hierarchy based on progressive cognitive and perceptual demands.

**Level 1: Homogeneous Counting.** This foundational tier focuses on basic visual grounding, requiring models to count a single type of object, action, or event under minimal attribute constraints.

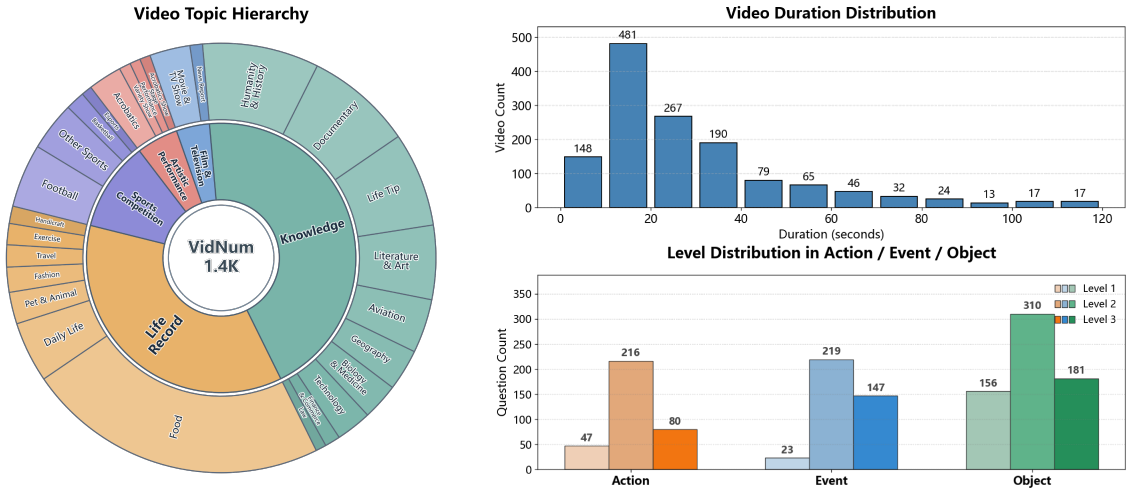


**Figure 2: Statistics of the *VidNum-1.4K*benchmark. Left: distribution of video topics. Top-right: distribution of video durations. Bottom-right: distribution across question levels and categories.**

Despite its straightforward premise, Level 1 rigidly tests a model's capacity for spatial-temporal tracking and object permanence. For example, in the "green wooden doors" scenario (Figure 1, top), the model must maintain a reliable "internal counter" as the camera pans, successfully disambiguating whether a door appearing in a new frame is a novel entity or merely a re-observation of a previous one. Unlike static image counting, this level establishes a strict baseline for quantifying homogeneous instances within continuous motion.

**Level 2: Constrained and Heterogeneous Counting.** The intermediate tier escalates complexity by shifting from homogeneous targets to multiple entity types or single entities defined by multi-attribute constraints. This level demands fine-grained discrimination beyond basic class-level detection. In the "dog counting" demonstration (Figure 1, middle), the model must differentiate various breeds based on heterogeneous features and aggregate their counts accurately. Furthermore, as these constrained scenarios frequently incorporate multi-shot videos, models are forced to perform robust re-identification to maintain numerical consistency across temporal gaps. For instance, the model must recognize that a brown-and-white dog appearing at both the start and end of the clip is the same individual, while ensuring that a Malinois-type dog seen from varying camera angles is not erroneously over-counted.

**Level 3: Compositional Numerical Reasoning.** The most advanced tier moves beyond direct perception to evaluate high-order cognitive capabilities, including comparison, calculation, estimation, and sequential ordering. These questions demand that models execute multi-step arithmetic operations strictly grounded in temporal evidence. As exemplified by the "soccer jersey" case (Figure 1, bottom), a model must first perform precise temporal localization to isolate the time windows of the first and second camera shots. Subsequently, it must count the players involved under strict jersey-color constraints, and finally execute a logical comparison between these two isolated events to deduce the correct answer. By demanding this chain of cross-modal, temporally-grounded logic, Level 3 pushes VLMs away from superficial language priors and rigorously assesses their capacity for genuine intelligence.

## 3.1 Benchmark Construction

Figure 3 (Top) illustrates the rigorous construction pipeline of *VidNum-1.4K*. To ensure the highest data quality, we adopted a fully manual, multi-stage human annotation pipeline. A dedicated team of approximately 300 full-time annotators was divided into four strictly isolated groups (A, B, C, and D) to construct, verify, and audit the benchmark.

**Source Video Collection.** To establish a general-purpose testbed for universal world understanding, our video collection team curated a highly heterogeneous visual corpus from a vast global pool. The raw queston set is systematically organized into a comprehensive two-tier topic hierarchy. As depicted in the topic distribution, the corpus spans five primary macro-categories: *Knowledge*, *Life Record*, *Sports Competition*, *Artistic Performance*, and *Film & Television*. Within these overarching domains, the videos cover a broad spectrum of fine-grained scenarios. These range from ubiquitous, high-frequency topics such as Food, Daily Life, and Humanity & History, to highly specialized fields requiring domain-specific visual grounding, such as Aviation, Biology & Medicine, and Acrobatics.

**Creation and Primary Verification (Groups A & B).** The initial question generation was executed by Group A (Creators). Annotators in this group were tasked with selecting specific video segments, defining precise timestamps, and designing challenging numerical reasoning questions based strictly on the visual evidence within that temporal window. Crucially, to proactively prevent VLMs from shortcutting the reasoning process via memorized language or world-knowledge priors, we enforced a strict "visual-description-only" policy for entity referencing. Annotators were explicitly prohibited from using proper nouns, celebrity names, or specific team affiliations in the question stems. For instance, as illustrated in the soccer jersey example (Figure 1, bottom row), instead of naming "Cristiano Ronaldo," the question relies entirely on objective visual attributes, such as "the short-haired man in the

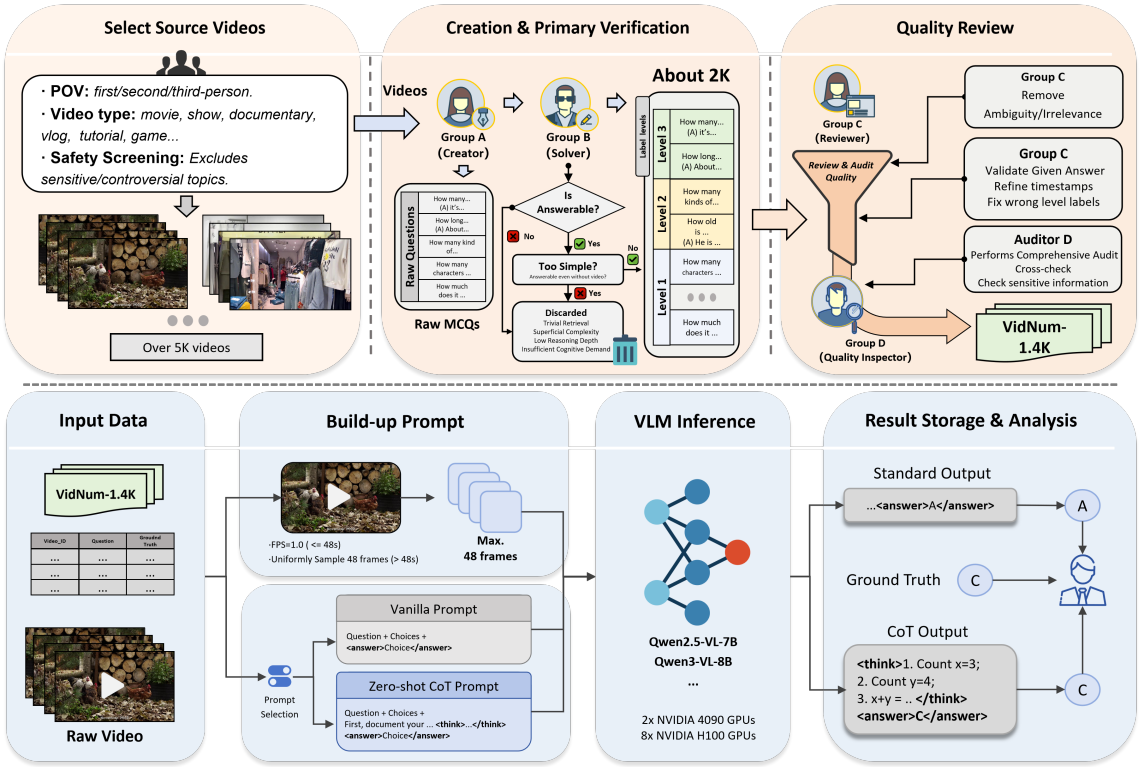


**Figure 3: Top: the data collection and annotation pipeline for *VidNum-1.4K*. Bottom: the evaluation pipeline on *VidNum-1.4K*.**

red jersey." Subsequently, Group B (Solvers) acted as the first layer of filtration by attempting to answer these generated questions. To prevent self-confirmation bias, Groups A and B consisted of strictly disjoint personnel. Group B discarded a question if: (1) it lacked a definitively correct answer, or (2) it was trivially simple and could be answered using pure language priors without watching the video. For all valid questions, Group B assigned a difficulty level (Level 1 to 3) based on our hierarchical definitions. This primary phase yielded an initial pool of approximately 2,000 candidate questions.

**Quality Review and Final Audit (Groups C & D).** To transform these candidates into a "Gold Standard" benchmark, the data entered a rigorous two-step verification phase. Group C (Reviewers) conducted a holistic review of the initial pool. Their primary objectives were to validate the absolute correctness of the answers, refine the timestamps to eliminate redundant video context that did not contribute to the reasoning process, and remove any ambiguous or irrelevant questions that deviated from the core theme of numerical reasoning. Finally, Group D (Quality Inspectors) executed the ultimate comprehensive double-check of all data validated by Group C, ensuring zero tolerance for annotation errors. We also ensured that personnel in Groups C and D were completely independent to prevent any downstream verification bias. Through this exhaustive human-in-the-loop pipeline, we finalized the VidNum-1.4K benchmark, yielding 1,379 meticulously verified, high-quality video-question pairs.

# 4 Experiments

## 4.1 Evaluation Models and Prompting Protocol

As illustrated in Figure 3 (Bottom), we evaluate a comprehensive suite of representative state-of-the-art (SOTA) open-source and closed-source VLMs on VidNum-1.4K. To rigorously assess their capabilities, we implement two distinct evaluation protocols: **Direct Answer** and **Zero-shot Chain-of-Thought (CoT)**. Across all settings, models are instructed to enclose their final choices within `<answer>` tags to facilitate automated parsing[1]. Under the CoT protocol, models are further prompted to articulate intermediate observations and calculation steps within `<think>` tags before committing to an option, effectively preventing the performance degradation caused by forced immediate responses. For visual input, we adopt a standardized frame sampling strategy to balance temporal resolution and computational efficiency: videos shorter than 48 seconds are sampled at 1 frame per second (FPS), while exactly 48 frames are uniformly extracted from longer sequences.

## 4.2 Main Results on VidNum1.4K

Table 1 summarizes the comprehensive evaluation results on the VidNum-1.4K benchmark, reporting accuracy (%) in a *NoCoT/CoT* format.

Overall, the results reveal that robust numerical video reasoning remains a significant challenge for current VLMs. Even the most

[1]A minor prompt adjustment is applied to *LLaVA-NeXT-7B-hf* to ensure formatting stability.

**Table 1: Main results on VidNum-1.4K across hierarchical levels and categories. Values represent *NoCoT/CoT* accuracy (%). A: Action, O: Object, E: Event. Within the open-source block, column-wise maxima are marked with bold+underline.**

| Model | Level 1 | | | Level 2 | | | Level 3 | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | Action | Object | Event | Action | Object | Event | Action | Object | Event | Avg. |
| **Open-source models** | | | | | | | | | | |
| InternVL2.5-8B | 27.66/29.79 | 32.69/33.77 | 26.09/47.83 | **35.65**/**36.28** | 38.39/38.83 | 39.73/41.10 | 36.25/33.75 | 38.67/48.30 | 45.58/48.98 | 37.64/40.07 |
| InternVL3-8B | 40.43/**44.68** | 30.77/30.77 | 30.43/39.13 | **35.65**/35.65 | 39.03/40.65 | 40.64/44.75 | 36.25/32.50 | 34.81/50.28 | 42.18/55.78 | 37.35/41.91 |
| InternVL3.5-8B | 38.30/34.04 | 32.69/33.55 | 21.74/39.13 | 32.41/33.80 | 32.58/40.32 | 34.25/43.84 | 33.75/36.25 | 45.30/40.88 | 40.14/55.78 | 35.39/40.35 |
| LLaVA-NeXT-7B | 27.66/27.03 | 28.21/25.78 | 21.74/29.41 | 20.56/21.94 | 25.48/22.78 | 22.02/25.00 | 21.25/21.67 | 23.33/25.17 | 26.21/24.44 | 24.03/24.05 |
| Qwen2.5-VL-7B | 34.04/40.43 | 28.21/30.72 | 34.78/39.13 | 32.41/25.35 | 33.87/37.86 | 38.81/31.05 | 37.50/27.50 | 33.15/38.33 | 38.78/44.76 | 34.45/34.31 |
| Qwen3-VL-8B | **42.55**/36.59 | 25.00/29.93 | 30.43/36.36 | 28.70/25.50 | 26.77/27.34 | 33.33/32.67 | **40.00**/33.82 | 37.57/41.88 | 36.05/39.83 | 31.69/32.08 |
| InternVL3-14B | 27.66/34.04 | 36.54/37.09 | 30.43/**52.17** | 32.41/32.87 | 38.71/44.41 | 43.38/46.33 | 33.75/**41.25** | 44.75/54.14 | 55.10/59.59 | 39.96/44.58 |
| InternVL3-38B | 36.17/38.30 | 33.33/39.10 | **56.52**/47.83 | 32.41/34.26 | 44.19/42.90 | **44.29**/**52.05** | 33.75/37.50 | 50.83/58.56 | **55.78**/56.46 | **42.57**/45.69 |
| InternVL3-78B | 31.91/40.43 | **40.38**/**40.38** | 47.83/43.48 | 30.56/34.72 | **44.52**/**45.48** | 41.55/45.21 | 38.75/38.75 | **51.38**/**59.12** | 51.02/**64.63** | 42.28/**46.41** |
| **Closed-source models** | | | | | | | | | | |
| Gemini-3-Flash | **51.06**/46.81 | 48.72/48.08 | 60.87/60.87 | 39.53/36.74 | 54.84/56.31 | 60.27/61.64 | 42.50/40.00 | 66.30/69.06 | **85.03**/76.03 | 56.60/55.74 |
| Gemini-3.1-Pro | 44.68/44.68 | **51.92**/**48.72** | **60.87**/**65.22** | **42.33**/**45.58** | **56.13**/**60.32** | **63.93**/**66.67** | **45.00**/**42.50** | **67.96**/**71.27** | 80.95/**85.03** | **57.98**/**60.30** |

advanced closed-source model, Gemini-3.1-Pro, barely reaches a 60% overall accuracy under the CoT setting, while representative open-source families (such as the InternVL series) typically struggle in the 25%–45% range. This stark performance gap confirms that VidNum-1.4K poses a substantial challenge to the physical world logic of state-of-the-art models.

A deeper analysis highlights three key bottlenecks in current architectures. First, CoT prompting serves as a critical catalyst for high-level compositional numerical reasoning (Level 3), indicating that while models may possess raw perceptual grounding, they often falter in logical aggregation without explicit step-by-step guidance. Second, across almost all evaluated models, action-based counting (A) consistently yields lower scores than object (O) or event (E) counting. This suggests that current VLMs struggle to track the continuous, fluid boundaries of dynamic actions compared to more static entities. Finally, the inclusion of multi-shot videos exposes a severe lack of "numerical consistency," as models frequently lose count or double-count across scene cuts. Collectively, these findings validate VidNum-1.4K as a demanding diagnostic testbed that exposes the fragility of current numerical video intelligence.

# 5 Discussion

## 5.1 Zero-shot CoT: A Double Edged Sword

Figure 4 reveals a nuanced but important pattern. At the aggregate level, introducing Zero-shot CoT tends to improve final accuracy, indicating that explicit intermediate reasoning can help VLMs resolve harder compositional dependencies in video counting. However, this improvement comes with a non-trivial cost: for every tested model, part of the gain is accompanied by a set of samples that were correct under direct answering but become incorrect after CoT prompting.

This "gain-with-regression" behavior suggests that current VLM competence is still fragile. In particular, some previously correct No-CoT predictions are likely not the result of stable causal reasoning, but of shortcut matching to statistical regularities, language priors, or memorized correlations from pretraining. When the model is forced to externalize a reasoning trace, these shortcuts are partially disrupted; the model can recover some genuinely hard cases, but it can also lose easy-but-unstable wins. The level/target breakdown in Figure 4 is consistent with this interpretation: CoT is most helpful where multi-step integration is required (especially higher-level/event-centric questions), yet it can introduce extra reasoning noise for lower-level perceptual counting. Overall, the evidence supports a cautious conclusion: today's VLMs have not yet learned a fully reliable world-grounded counting mechanism, and still rely substantially on heuristic "intuition" rather than robust understanding of dynamic physical scenes.

## 5.2 What's happening while scaling up?

To understand how parameter expansion influences VLMs' numerical capabilities, we analyze the scaling behavior of the open-source InternVL3 series (ranging from 8B to 78B parameters). As illustrated in Figure **??**, increasing model capacity yields a general upward trend in performance, though the gains are highly uneven across different task levels. Notably, Level 3 (Compositional Numerical Reasoning) exhibits the most substantial improvements, particularly under the Zeroshot CoT setting, where accuracy climbs from 48.8% at the 8B scale to 57.1% at 78B. This suggests that scaling up parameter count effectively enhances a model's ability to perform high-level logical deductions based on visual evidence. Conversely, Level 1 shows only steady but modest improvements, while Level 2 remains surprisingly stagnant across all model sizes in both settings. This stagnation indicates that fine-grained instance tracking and robust cross-shot re-identification (the core challenges of Levels 1 and 2) represent fundamental perceptual bottlenecks. Ultimately, these results demonstrate that while scaling improves high-level arithmetic reasoning, foundational visual grounding in dynamic

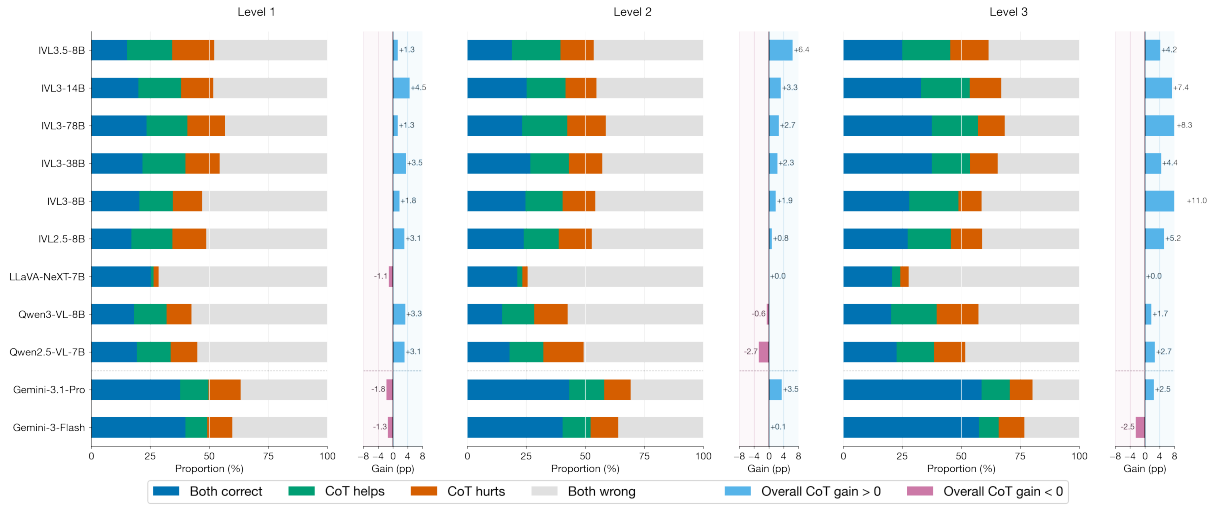


**Figure 4: Impact of Chain-of-Thought (CoT) prompting on open-source VLMs across diverse task dimensions. (a) Mean accuracy gain/loss (in percentage points) across the three hierarchical levels of VidNum-1.4K. (b) Mean accuracy gain/loss categorized by counting targets (object, action, and event). The results indicate that while CoT facilitates high-level reasoning in Level 3 and Event tasks, it tends to hinder performance in lower-level perceptual counting.**

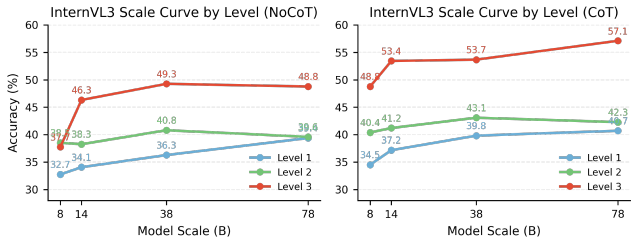


**Figure 5: Scaling trends of the InternVL3 series across VidNum-1.4K hierarchical levels. The left and right panels illustrate performance under NoCoT and CoT settings, respectively.**

scenes cannot be easily resolved by simply increasing parameter capacity alone.

## 6 Conclusion

We introduce *VidNum-1.4K*, a hierarchical and multi-domain benchmark for diagnosing numerical video intelligence in Vision-Language Models. Extensive experiments show a clear performance ceiling: even frontier models struggle with temporal counting and compositional numerical reasoning. The consistent reasoning gap across model families suggests that current VLMs still rely on statistical shortcuts rather than robust, world-grounded counting mechanisms. By combining fragmented temporal contexts with multi-step reasoning requirements, *VidNum-1.4K*provides a strong stress test for exposing these weaknesses. We hope this benchmark will support the development of next-generation models with more reliable and physically grounded video understanding.

## 7 Code and Data Availability

To support reproducibility and facilitate future research in video-based numerical reasoning, we make our benchmark fully publicly available. The complete VidNum-1.4K benchmark, including video metadata, multiple-choice question annotations, and evaluation splits, can be accessed through our project page at https://VidNumTeam.github.io. Furthermore, the source code for data preprocessing, evaluation protocols, and baseline model inference is hosted on GitHub at https://github.com/VidNumTeam/VidNum.git. The source code is released under the **MIT License**, and the benchmark annotations are released under the **CC BY 4.0 License**.

## References

[1] Manoja S. Acharya, Chih-Hui Hsieh, Yezhou Yang, R. K. K. Rao, et al. 2018. TallyQA: Answering Complex Counting Questions. *arXiv preprint arXiv:1810.12440* (2018).

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Yana Hasson, et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198* (2022).

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631* (2025).

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[6] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2020. Counting Out Time: Class Agnostic Video Repetition Counting in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[9] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. *arXiv preprint arXiv:1804.04527* (2018).

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Yunseok Jang, Yale Song, Chris D. Manning, Bhavani Thuraisingham, and Jiawei Han. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The Kinetics Human Action Video Dataset. In *arXiv preprint arXiv:1705.06950*.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597* (2023).

[16] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485* (2023).

[18] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424* (2023).

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020* (2021).

[20] Tom F. H. Runia, Cees G. M. Snoek, and Arnold W. M. Smeulders. 2018. Real-World Repetition Estimation by Div, Grad and Curl. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[22] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[24] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next Phase of Question Answering to Explain Temporal Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

[27] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106* (2025).

[28] Hang Zhang, Xin Li, Lidong Bing, et al. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858* (2023).

[29] Honglei Zhang, Xiang Tao, Shanshe Wang, Wen Gao, and Qi Tian. 2020. Context-Aware and Scale-Insensitive Temporal Repetition Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.