Qwen3-VL's 256K-Token Context Window in Long-Video Understanding on Video-MME

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: What is the impact of Qwen3-VL's 256K token context window on long-video understanding tasks compared to dense variants of similar parameter counts on the Video-MME benchmark. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Qwen3-VL Technical Report. Research question: What is the impact of Qwen3-VL's 256K token context window on long-video understanding tasks compared to dense variants of similar parameter counts on the Video-MME benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 6 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

6 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Qwen3-VL-235B-A22B-Thinking achieves the highest score of 78.7 on MMStar.	×	0.05
Gemini-2.5-Pro's Thinking mode delivers the best overall performance on MMStar.	×	0.04
Qwen3-VL-235B-A22B-Instruct obtains the highest scores on MMBench and RealWorldQA, with 89.3/88.9 and 79.2, respectively	×	0.04
Qwen3-VL-32B-Thinking achieves the highest scores on MMBench and RealWorldQA, with 89.5/89.5 and 79.4, respectively.	×	0.03
Qwen3-VL-32B-Instruct outperforms the Thinking variant on RealWorldQA, scoring 79.0.	×	0.04
Qwen3-VL-8B achieves the highest performance across all five benchmarks.	×	0.08
On MMBench-EN, the score in 'thinking' mode increases from 79.9 for the 2B model to 85.3 for the 8B model.	×	0.03
On MMStar, the score rises from 68.1 (2B, thinking) to 75.3 (8B, thinking).	×	0.02
Qwen3-VL-235B-A22B-Instruct achieves the best reported results among non-thinking or low-thinking-budget models on multi	×	0.04
Qwen3-VL-235B-A22B-Thinking achieves state-of-the-art results on MathVistamini, MathVision, MathVersemini, ZeroBench, Lo	×	0.05
Qwen3-VL-32B demonstrates significant advantages on MMMU, MMMU-Pro, MathVistamini, MathVision, MathVisionWP, We-Math, Ma	×	0.05

References

- http://arxiv.org/abs/2511.21631v2

- http://arxiv.org/abs/2604.05015v1
- http://arxiv.org/abs/2405.21075v3