Gemini 1.5 Flash and Pro Zero-Shot Cross-Domain Performance on MMBench

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How do Gemini 1.5 Flash and Pro perform in zero-shot cross-domain adaptation tasks on the MMBench benchmark, and what are the trade-offs in accuracy and inference time between the two models. 12 claims were extracted from source literature; 11 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.4/10. 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: How do Gemini 1.5 Flash and Pro perform in zero-shot cross-domain adaptation tasks on the MMBench benchmark, and what are the trade-offs in accuracy and inference time between the two models?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

14 papers retrieved. 12 claims extracted; 11 independently verified. Quality review score: 7.4/10.

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 is the most capable vision-language model in the Qwen series to date.	✓	0.22
Qwen3-VL natively supports interleaved contexts of up to 256K tokens.	✓	0.22
Qwen3-VL seamlessly integrates text, images, and video.	✓	0.15
The Qwen3-VL model family includes dense variants with parameter counts of 2B, 4B, 8B, and 32B.	✓	0.17
The Qwen3-VL model family includes mixture-of-experts variants with configurations 30B-A3B and 235B-A22B.	✓	0.18
Qwen3-VL demonstrates stronger pure-text understanding than comparable text-only backbones in several cases.	✓	0.22
Qwen3-VL enables faithful retention, retrieval, and cross-referencing across long documents and videos via a native 256K	✓	0.27
Qwen3-VL demonstrates leading performance on the MMMU benchmark.	×	0.09
Qwen3-VL demonstrates leading performance on visual-math benchmarks including MathVista and MathVision.	✓	0.15
Qwen3-VL architecture includes an enhanced interleaved-MRoPE for stronger spatial-temporal modeling across images and vi	✓	0.26
Qwen3-VL architecture includes DeepStack integration to leverage multi-level ViT features for tighter vision-language al	✓	0.18
Qwen3-VL utilizes text-based time alignment for video, evolving from T-RoPE to explicit textual timestamp alignment.	✓	0.27

References

- https://openalex.org/W7120272020
- https://doi.org/10.1007/s44267-025-00099-6
- https://doi.org/10.48550/arxiv.2511.21631