

Zen3-VL: A Packaging of Qwen3-VL-30B-A3B

Technical Note v2025.06

Zen LM Research Team
research@zenlm.org
papers.zenlm.org

June 2025

Abstract

Zen3-VL is *not* a from-scratch model. It is a redistribution and packaging of **Qwen3-VL-30B-A3B**, the Mixture-of-Experts vision-language model developed by the Qwen team at Alibaba Cloud and released under the Apache-2.0 license [1]. The upstream model uses the `Qwen3VLMoeForConditionalGeneration` architecture and is a roughly 31B-total-parameter MoE with about 3.3B activated parameters per token — not the 72B dense flagship described in earlier internal drafts. This note documents the upstream architecture as published by its authors, records the provenance and license obligations, and describes the thin packaging layer (weight conversion, quantization, serving configuration) Zen LM applies on top of the upstream checkpoint. No architectural novelty, no separate pre-training run, and no independently produced benchmark results are claimed. For authoritative capability numbers, see the upstream model card and the Qwen3-VL Technical Report [1, 2].

Contents

1	Provenance and Scope	2
2	Upstream Architecture (as published by Qwen)	2
2.1	Three-Module Structure	2
2.2	Named Architectural Features	2
2.3	Context Length and Resolution	3
2.4	Items Removed From Earlier Drafts	3
3	Packaging and Redistribution	3
3.1	Conversion and Quantization	3
3.2	Attribution Requirements	3
4	Evaluation: Use Upstream Numbers	3
5	Related Work	4
6	Conclusion	4

1 Provenance and Scope

Earlier internal drafts of this document described a homegrown “Zen MoDE (Mixture of Distilled Experts)” backbone at 72B parameters, a bespoke document-specialized vision encoder, a “scientific figure parser,” and a table of benchmark scores presented as in-house results. That framing was inaccurate and has been removed. The artifact distributed as “Zen3-VL” is the Qwen3-VL-30B-A3B checkpoint published by Alibaba Cloud’s Qwen team [1]. This note exists to attribute the work correctly and to make the redistribution terms explicit.

Upstream model. Qwen3-VL-30B-A3B is a Mixture-of-Experts vision-language model. Per the model card, it has approximately 31B total parameters with about 3.3B activated per token, uses the `Qwen3VLMoeForConditionalGeneration` class, and supports a native 256K-token context expandable to 1M tokens [1]. It is published in both Instruct and reasoning-oriented “Thinking” editions.

License. The upstream weights are released under **Apache-2.0**. This is permissive and allows commercial use and redistribution, but a redistributor must preserve copyright and license notices, include the upstream NOTICE file if present, and state any modifications. Any Zen LM redistribution must ship the upstream LICENSE and attribution intact and must not misrepresent the model’s origin.

Correction of parameter scale. The “72B” figure and the “128 experts, top-6, 660B effective parameters” configuration in earlier drafts were fabricated and do not describe this model. The upstream Qwen3-VL-30B-A3B is a ~31B-total MoE with ~3.3B activated parameters. The precise per-layer expert count and routing top- k are defined by the upstream configuration files; this note does not restate or invent them.

What “packaging” means here. Zen LM does not retrain or re-architect the model. The packaging layer is limited to weight-format conversion, optional post-training quantization, and serving configuration. These steps do not change the model’s identity or its measured capabilities.

2 Upstream Architecture (as published by Qwen)

The description below summarizes the architecture as documented by the upstream authors in the Qwen3-VL model card and Technical Report [1, 2]. It is reproduced for convenience and is *not* a Zen LM contribution.

2.1 Three-Module Structure

Following the Qwen2.5-VL design, Qwen3-VL adopts a three-module structure:

- **Vision encoder** — a ViT-based encoder that processes images and video frames at high resolution.
- **Vision-language merger** — an MLP-based module that projects visual features into the language model’s embedding space.
- **LLM backbone** — the Mixture-of-Experts language model that performs multimodal reasoning and text generation.

2.2 Named Architectural Features

The upstream authors highlight several features of Qwen3-VL [1]:

- **Interleaved-MRoPE** — multimodal rotary position embeddings with full-frequency allocation over the time, width, and height axes, intended to improve long-horizon video reasoning.

- **DeepStack** — fusion of multi-level ViT features to capture fine-grained visual detail and sharpen image–text alignment.
- **Text–Timestamp Alignment** — a video temporal-modeling mechanism that moves beyond T-RoPE toward timestamp-grounded event localization.

These are upstream Qwen3-VL contributions and are documented by their authors [1, 2].

2.3 Context Length and Resolution

Per the model card, the native context window is 256K tokens, expandable to 1M tokens, and the vision stack supports high-resolution image input with adaptive handling of varying aspect ratios [1]. Exact resolution and tiling parameters are defined upstream.

2.4 Items Removed From Earlier Drafts

The following claims appeared in prior drafts and are **withdrawn** because they were not substantiated and were presented as in-house work: a 72B “Zen MoDE” backbone; a specific expert-cluster specialization map (“experts 1–24 handle text,” etc.); a 6.5B-parameter from-scratch document vision encoder with stated training scale; a “scientific figure parser” head with a per-chart-type extraction-accuracy table; a “chart-to-data synthetic pipeline” producing 100M samples; and all benchmark tables comparing a fictional “Zen-VL / Zen2-VL / Zen3-VL” lineage. Any genuine capabilities in these areas belong to the upstream Qwen3-VL model.

3 Packaging and Redistribution

3.1 Conversion and Quantization

Zen LM converts the published Qwen3-VL-30B-A3B weights to the internal serving format and may apply post-training quantization for deployment. These transforms are mechanical; they do not constitute a new training run and do not produce a new model. Where quantization is applied, the scheme and bit-width should be recorded alongside the artifact so downstream users understand any quality trade-off relative to the upstream checkpoint.

3.2 Attribution Requirements

Because the upstream license is Apache-2.0, every redistributed Zen3-VL artifact must:

1. Retain the upstream copyright and Apache-2.0 LICENSE text.
2. Include the upstream NOTICE file, if any, and not remove attribution.
3. State plainly that the model is a packaging of Qwen3-VL-30B-A3B by the Qwen team at Alibaba Cloud, and describe any modifications (e.g. quantization).
4. Avoid any naming or marketing that implies the architecture or weights are original Zen LM research.

4 Evaluation: Use Upstream Numbers

This note deliberately reports **no** Zen-produced benchmark scores. The Qwen team publishes Qwen3-VL evaluations on standard vision-language suites — including general multimodal reasoning (e.g. MMMU, MMBench), document and OCR understanding (e.g. DocVQA, OCRBench,

InfoVQA), chart and structured-data tasks (e.g. ChartQA), and mathematical visual reasoning (e.g. MathVista) — in the official model card and Technical Report [1, 2].

Guidance for downstream reporting. Cite the upstream-reported numbers with explicit attribution to Qwen, and link to the model card. If Zen LM applies quantization, any re-measured scores should be reported as “Qwen3-VL-30B-A3B (Zen packaging, N -bit)” with the evaluation harness named, framed as a measurement of the upstream model under a deployment configuration — never as the result of a distinct model.

5 Related Work

The relevant prior art is the upstream lineage itself: the Qwen-VL and Qwen2.5-VL line of vision-language models from Alibaba Cloud, of which Qwen3-VL is the current generation [1, 2]. The contrastive vision-language pre-training paradigm [3] and end-to-end document understanding [4] are foundational to the broader field; the specific Interleaved-MRoPE, DeepStack, and Text-Timestamp Alignment mechanisms used here are Qwen3-VL contributions.

6 Conclusion

“Zen3-VL” denotes a packaged redistribution of Alibaba Cloud’s Apache-2.0 Qwen3-VL-30B-A3B (an ~ 31 B-total / ~ 3.3 B-active MoE), not an independent 72B model. The honest description of the work is: convert, optionally quantize, serve, and attribute. All architectural credit and all capability claims belong to the upstream Qwen3-VL authors and should be cited to them.

Acknowledgments

We acknowledge the Qwen team at Alibaba Cloud as the authors of Qwen3-VL-30B-A3B, the model packaged and redistributed here under the Apache-2.0 license.

References

- [1] Qwen Team, Alibaba Cloud, “Qwen3-VL-30B-A3B-Instruct,” Hugging Face model card, 2025. <https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct>
- [2] Qwen Team, Alibaba Cloud, “Qwen3-VL Technical Report,” arXiv:2511.21631, 2025. <https://arxiv.org/abs/2511.21631>
- [3] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” *ICML*, 2021.
- [4] Y. Xu et al., “LayoutLM: Pre-Training of Text and Layout for Document Image Understanding,” *KDD*, 2020.