

Zen-VL: A Packaging of the Qwen3-VL Series

Technical Note v2025.06

Zen LM Research Team
research@zenlm.org

June 2025

Abstract

Zen-VL is *not* a from-scratch model. It is a redistribution and packaging of the **Qwen3-VL** vision-language model series developed by the Qwen team at Alibaba Cloud and released under the Apache-2.0 license [1]. The Zen-VL packaging targets the compact dense editions of that series — **Qwen3-VL-4B** and **Qwen3-VL-8B** (each available in Instruct and reasoning-oriented “Thinking” editions) — with the larger Mixture-of-Experts variants (30B-A3B and 235B-A22B) available upstream for higher-capacity deployments. 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. No architectural novelty, no separate pre-training run, and no independently produced benchmark results are claimed. For authoritative capability numbers, see the upstream model cards and the Qwen3-VL Technical Report [1, 2, 3].

Contents

1	Provenance and Scope	2
2	Upstream Architecture (as published by Qwen)	2
2.1	Model Family	2
2.2	Three-Module Structure	2
2.3	Named Architectural Features	3
2.4	Context and Resolution	3
2.5	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	4
5	Related Work	4
6	Limitations	4
7	Conclusion	4

1 Provenance and Scope

Earlier internal drafts of this document described a homegrown 7B “Zen MoDE (Mixture of Distilled Experts)” backbone with a ViT-L/14 encoder, a bespoke dynamic-tiling scheme, a two-stage training recipe, and a set of benchmark tables presented as in-house results. That framing was inaccurate and has been removed. The artifact distributed as “Zen-VL” is a packaging of the Qwen3-VL series published by Alibaba Cloud’s Qwen team [1]. This note exists to attribute the work correctly and to make the redistribution terms explicit.

Upstream models. Qwen3-VL is the current-generation multimodal LLM series from the Qwen team [3]. It spans dense models at 4B and 8B parameters and MoE variants at 30B-A3B and 235B-A22B, each offered in Instruct and “Thinking” editions [2]. Zen-VL packages the dense **4B** and **8B** editions as the compact entry points; the upstream MoE editions are documented separately and may be packaged under the Zen3-VL note.

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 model identity. The “7B Zen MoDE” backbone, the “307M ViT-L/14” encoder, the specific tiling configuration table, the ten-million-pair Stage-1 and two-million-pair Stage-2 training recipe, and all “Zen-VL vs. Comp. A/B/C” benchmark tables in earlier drafts were fabricated and do not describe these models. The actual artifacts are the Qwen3-VL-4B and Qwen3-VL-8B dense checkpoints.

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.

2 Upstream Architecture (as published by Qwen)

The description below summarizes the architecture as documented by the upstream authors in the Qwen3-VL model cards and Technical Report [2, 3]. It is reproduced for convenience and is *not* a Zen LM contribution.

2.1 Model Family

Table 1: Qwen3-VL series as published upstream (Alibaba Cloud, Apache-2.0). Zen-VL packages the dense 4B/8B editions.

Upstream model	Type	Editions
Qwen3-VL-4B	Dense	Instruct, Thinking
Qwen3-VL-8B	Dense	Instruct, Thinking
Qwen3-VL-30B-A3B	MoE	Instruct, Thinking
Qwen3-VL-235B-A22B	MoE	Instruct, Thinking

2.2 Three-Module Structure

Following the Qwen2.5-VL design, each Qwen3-VL model adopts a three-module structure: a ViT-based vision encoder, an MLP-based vision-language merger that projects visual features into the language embedding space, and the language model backbone (dense for the 4B/8B editions). Exact layer counts, hidden sizes, and patch configurations are defined by the upstream configuration files and are not restated or invented here.

2.3 Named Architectural Features

The upstream authors highlight several features shared across the Qwen3-VL series [2]:

- **Interleaved-MRoPE** — multimodal rotary position embeddings with full-frequency allocation over the time, width, and height axes.
- **DeepStack** — fusion of multi-level ViT features to capture fine-grained visual detail and improve image-text alignment.
- **Text-Timestamp Alignment** — timestamp-grounded event localization for stronger video temporal modeling.

2.4 Context and Resolution

Per the upstream model cards, the Qwen3-VL series provides a native 256K-token context window, expandable to 1M tokens, with high-resolution image and video input [2]. Multi-image conversations and video understanding are supported natively. Exact resolution handling is defined upstream.

2.5 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: the 7B “Zen MoDE” backbone and its 7.5B-total component table; the ViT-L/14 (307M) encoder spec and a 1024-token-per-tile patching claim; the specific dynamic-tiling configuration and visual-token-count tables; the Stage-1/Stage-2 training recipe with stated optimizer, batch size, and data counts; the SFT data-composition table; and *all* evaluation tables (multimodal benchmarks, document understanding, high-resolution ablation, multi-image, and hallucination). These numbers were invented. Genuine capabilities belong to the upstream Qwen3-VL models.

3 Packaging and Redistribution

3.1 Conversion and Quantization

Zen LM converts the published Qwen3-VL-4B and Qwen3-VL-8B weights to the internal serving format and may apply post-training quantization for deployment, which is especially relevant for the edge-oriented 4B edition. 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 Zen-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 (4B/8B) 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 diagram tasks (e.g. ChartQA, AI2D), and object-hallucination evaluation (e.g. POPE, HallusionBench) — in the official model cards and Technical Report [2, 3].

Guidance for downstream reporting. Cite upstream-reported numbers with explicit attribution to Qwen, and link to the model card for the specific size (4B or 8B). If Zen LM applies quantization, any re-measured scores should be reported as “Qwen3-VL-4B/8B (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 [2, 3]. Contrastive vision-language pre-training [4] and projector-based connection of frozen vision encoders to language models (BLIP-2 [5], LLaVA [6]) are foundational to the broader field; the Interleaved-MRoPE, DeepStack, and Text–Timestamp Alignment mechanisms used here are Qwen3-VL contributions.

6 Limitations

The limitations of Zen-VL are the limitations of the upstream Qwen3-VL models it packages, as documented by their authors. Additionally, any post-training quantization applied during packaging may reduce quality relative to the upstream full-precision checkpoints; the degree depends on the chosen scheme and bit-width and should be measured and disclosed per artifact.

7 Conclusion

“Zen-VL” denotes a packaged redistribution of Alibaba Cloud’s Apache-2.0 Qwen3-VL dense models (4B and 8B), not an independent 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 the Qwen3-VL series, the models packaged and redistributed here under the Apache-2.0 license.

References

- [1] Qwen Team, Alibaba Cloud, “Qwen3-VL,” GitHub repository, 2025. <https://github.com/QwenLM/Qwen3-VL>
- [2] Qwen Team, Alibaba Cloud, “Qwen3-VL-4B-Instruct,” Hugging Face model card, 2025. <https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct>
- [3] Qwen Team, Alibaba Cloud, “Qwen3-VL Technical Report,” arXiv:2511.21631, 2025. <https://arxiv.org/abs/2511.21631>

- [4] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” *ICML*, 2021.
- [5] J. Li et al., “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” *ICML*, 2023.
- [6] H. Liu et al., “Visual Instruction Tuning,” *NeurIPS*, 2023.
- [7] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR*, 2021.