

Zen AI Model Family

Zen-Designer-235B-A22B-Thinking

Vision-Language Reasoning for Design

Technical Whitepaper v1.0

Zach Kelling*
research@hanzo.ai

Zoo Labs Foundation
foundation@zoolabs.org

September 2025

Abstract

We present **Zen-Designer-235B-A22B-Thinking**, a deliberative vision-language model for design reasoning and visual analysis. The model is a derivative of the open-weight **Qwen3-VL-235B-A22B** mixture-of-experts foundation (Apache 2.0), adapted by Hanzo AI and the Zoo Labs Foundation to emit explicit intermediate reasoning before producing a final answer on visual and layout tasks. It is a sparse MoE with 235B total parameters of which approximately 22B are active per token, served under the `Qwen3VLForConditionalGeneration` architecture. Weights are released openly under Apache 2.0 at huggingface.co/zenlm/zen-designer-235b-a22b-thinking. This whitepaper documents the architecture as published in the released configuration and the intended use of the thinking variant; it does not claim novel pretraining or report benchmark numbers that have not been independently reproduced.

Contents

1	Introduction	2
1.1	Provenance and Licensing	2
1.2	Scope	2
2	Architecture	2
2.1	Model Configuration	2
2.2	Mixture of Experts	3
2.3	Deliberative Reasoning	3
2.4	Vision-Language Integration	3
3	Intended Use	3
3.1	Primary Applications	3
3.2	Thinking vs. Instruct	3
3.3	Usage Example	3
4	Deployment	4
4.1	Available Formats	4
4.2	Hardware Footprint	4

*zach@lux.network

5	Limitations	4
6	Conclusion	4
A	Model Card	5

1 Introduction

Zen-Designer-235B-A22B-Thinking is the deliberative member of the Zen-Designer line. It is tuned to produce a chain of intermediate reasoning steps before a final response, trading latency for improved performance on multi-step visual-reasoning tasks. It is built on the open-weight **Qwen3-VL-235B-A22B** foundation released by the Qwen team under the Apache 2.0 license. A companion *instruct* variant ([zen-designer-235b-a22b-instruct](#)) provides lower-latency, direct responses.

1.1 Provenance and Licensing

This model is a derivative of an upstream open-weight foundation. We state the lineage explicitly to avoid any overclaiming:

- **Upstream foundation:** Qwen3-VL-235B-A22B (vision-language MoE), Apache 2.0.
- **Derivative:** Zen-Designer-235B-A22B-Thinking, reasoning-tuned for design tasks.
- **License:** Apache 2.0 (inherited; redistribution permitted with attribution).
- **Repository:** [zenlm/zen-designer-235b-a22b-thinking](#) on Hugging Face.

1.2 Scope

- **Sparse MoE efficiency:** ~ 22 B parameters active per token from a 235B total pool.
- **Deliberative reasoning:** explicit intermediate steps before a final answer.
- **Vision-language:** image-conditioned reasoning and text-conditioned design assistance.
- **Extended context:** 131K-token maximum position embeddings (per released config).

2 Architecture

2.1 Model Configuration

The values below are taken directly from the released model configuration (`config.json`); they describe the published architecture rather than any internally measured performance.

Component	Specification
Architecture	<code>Qwen3VLForConditionalGeneration</code>
Total Parameters	235B (sparse MoE)
Active Parameters	~ 22 B per token
Experts / Active per Token	64 / 4
Text Hidden Size	8192
Text Layers	80
Attention Heads / KV Heads	64 / 8
Vision Hidden Size	2048
Vision Layers	48
Vision Patch / Image Size	14 / 2048
Vocabulary Size	151,936
Max Position Embeddings	131,072
Precision	bfloat16

Table 1: Zen-Designer-235B-A22B-Thinking configuration (from released `config.json`).

2.2 Mixture of Experts

The text backbone is a sparsely activated mixture-of-experts transformer: each token is routed to a small subset (top-4 of 64) of expert feed-forward blocks per layer, so that only a fraction of the 235B total parameters is exercised on any given forward pass. This is the standard MoE efficiency trade-off described by Shazeer et al. [3] and Fedus et al. [4], inherited from the Qwen3-VL foundation.

2.3 Deliberative Reasoning

The thinking variant is tuned to emit an explicit reasoning trace prior to its final response, following the chain-of-thought paradigm [5]. This typically improves accuracy on multi-step visual and layout-reasoning tasks at the cost of additional generated tokens and latency relative to the instruct variant.

2.4 Vision-Language Integration

A vision encoder (48 layers, hidden size 2048, patch size 14) produces visual tokens that are consumed by the text decoder under the `Qwen3VLForConditionalGeneration` interface, enabling image-conditioned reasoning alongside text-only generation.

3 Intended Use

3.1 Primary Applications

- Multi-step UI/UX design analysis and critique
- Layout and information-architecture reasoning
- Visual question answering requiring step-by-step justification
- Design-system review with explicit rationale
- Accessibility evaluation with documented reasoning

3.2 Thinking vs. Instruct

The thinking variant is preferred where multi-step visual reasoning or self-verification adds value and additional latency is acceptable. For low-latency, direct responses, use the companion instruct variant.

3.3 Usage Example

```
1 from transformers import AutoModelForVision2Seq, AutoProcessor
2
3 model = AutoModelForVision2Seq.from_pretrained(
4     "zenlm/zen-designer-235b-a22b-thinking")
5 processor = AutoProcessor.from_pretrained(
6     "zenlm/zen-designer-235b-a22b-thinking")
7
8 # Image-conditioned reasoning
9 inputs = processor(images=image, text="Analyze this UI", return_tensors
10     ="pt")
11 outputs = model.generate(**inputs)
12 analysis = processor.decode(outputs[0])
```

4 Deployment

4.1 Available Formats

- **SafeTensors**: native bfloat16 weights (sharded).
- **GGUF** / **MLX**: community quantizations may be provided separately.

4.2 Hardware Footprint

Because the model carries 235B total parameters, full-precision serving requires multi-accelerator deployment; quantization reduces this footprint at some quality cost. The thinking variant additionally emits more tokens per query than the instruct variant, which increases end-to-end latency. We do not publish a single canonical latency figure, as throughput depends heavily on the serving stack, batch size, quantization, and hardware.

5 Limitations

- **Inherited behavior**: as a derivative of Qwen3-VL, the model inherits the capabilities, biases, and limitations of its upstream foundation and training data.
- **No independent benchmark claims**: this document intentionally omits benchmark scores that have not been independently reproduced for this derivative.
- **Latency**: the deliberative reasoning trace increases token count and latency relative to the instruct variant.
- **Resource intensity**: the 235B parameter count makes self-hosting costly relative to smaller members of the Zen family.
- **Visual generation**: the model assists with design *reasoning* and analysis; pixel-level image synthesis is out of scope for the language backbone.

6 Conclusion

Zen-Designer-235B-A22B-Thinking packages an open-weight Qwen3-VL-235B-A22B foundation as a deliberative design and visual-reasoning assistant, released openly under Apache 2.0. We have described its architecture as published and stated its provenance plainly so that downstream users can make informed decisions about licensing, capability, and cost.

Acknowledgments

We thank the Qwen team for releasing the Qwen3-VL foundation under Apache 2.0, and the open-source community and the teams at Hanzo AI and Zoo Labs Foundation for their contributions to this work.

References

- [1] Vaswani, A. et al. (2017). Attention Is All You Need. NeurIPS 2017.
- [2] Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.
- [3] Shazeer, N. et al. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. ICLR 2017.
- [4] Fedus, W. et al. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. JMLR 2022.
- [5] Wei, J. et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022.
- [6] Qwen Team (2025). Qwen3-VL: Vision-Language Models. huggingface.co/Qwen.

A Model Card

Field	Value
Model Name	Zen-Designer-235B-A22B-Thinking
Upstream Foundation	Qwen3-VL-235B-A22B (Apache 2.0)
Version	1.0.0
Release Date	September 2025
License	Apache 2.0
Repository	huggingface.co/zenlm/zen-designer-235b-a22b-thinking
Documentation	github.com/zenlm/zen
Contact	research@hanzo.ai

Table 2: Model Card Information