

Zen3-Omni: Third-Generation Unified Multimodal Intelligence

Technical Whitepaper v2025.06

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

June 2025

Abstract

Zen3-Omni advances multimodal AI by unifying text, image, audio, and video understanding in a single 7B parameter model, featuring a novel cross-modal attention architecture that enables fluent reasoning across modalities with state-of-the-art performance on multimodal benchmarks. The third generation introduces substantial improvements in video understanding through temporal sparse attention and real-time audio processing via streaming CTC decoders integrated natively into the Zen MoDE (Mixture of Distilled Experts) backbone. Zen3-Omni achieves MMBench 86.2%, AudioCaps CLAP 0.523, VideoQA 79.4%, and OCRBench 82.1%, making it the leading 7B-class multimodal model across all four primary modalities simultaneously. The model supports streaming inference for real-time audio/video inputs with end-to-end latency below 300ms on a single A100 GPU.

Contents

1	Introduction	3
2	Architecture	3
2.1	Unified Token Space	3
2.2	Cross-Modal Attention	3
2.3	Temporal Sparse Attention for Video	4
2.4	Zen MoDE Expert Routing with Modality Balance	4
2.5	Streaming CTC Audio Decoder	4
3	Training	4
3.1	Pre-Training Data	4
3.2	Training Curriculum	5
3.3	Distillation from Zen3-VL	5
4	Evaluation	5
4.1	Vision-Language	5
4.2	Audio Understanding	5
4.3	Video Understanding	5
4.4	Cross-Modal Reasoning	6
4.5	Streaming Latency	6

5	Related Work	6
5.1	Multimodal Language Models	6
5.2	Audio-Language Models	6
5.3	Video Understanding	6
6	Conclusion	6

1 Introduction

The dominant paradigm in multimodal AI has been modality-specialized towers: separate vision encoders, audio encoders, and video encoders, each pre-trained independently and fused into a shared language model backbone via adapter layers. While effective, this architecture introduces seams at modality boundaries — the model reasons about each modality separately rather than in a unified representational space.

Zen3-Omni is built on a fundamentally different premise: a single cross-modal attention mechanism processes tokens from all modalities together, allowing the model to form joint representations where a spoken word, its transcript, and a visual depiction of the concept co-attend to one another naturally.

This paper makes the following contributions:

1. **Unified cross-modal attention** — A single attention mechanism that handles text, image patch, audio frame, and video temporal tokens without modality-specific routing, enabling true multi-modal chain-of-thought reasoning.
2. **Temporal sparse attention for video** — A hierarchical temporal attention pattern that processes 60fps video at 1-second granularity without quadratic sequence length blowup, enabling up to 10-minute video inputs.
3. **Streaming CTC audio decoder** — Real-time audio transcription and understanding with 40ms chunk processing latency, enabling live speech input.
4. **Third-generation Zen MoDE scaling** — Expert routing updated with modality-aware load balancing that prevents expert collapse on underrepresented modalities.

2 Architecture

2.1 Unified Token Space

All four modalities are projected into a shared 4096-dimensional token space before entering the transformer backbone:

Text: Standard BPE tokenization with a 150,528-token vocabulary. Each token $t_i \in \mathbb{R}^{4096}$.

Image: ViT-style patchification at 14×14 pixel patches. High-resolution images use dynamic resolution tiling (up to 4×4 tiles at 448×448 base resolution). Each image produces 256–4096 patch tokens.

Audio: Mel spectrogram with 128 mel bins, 10ms hop, processed in 40ms chunks. A 3-layer CNN produces audio tokens at 40ms granularity. 30 seconds of audio yields 750 tokens.

Video: Frame sampling at 1 fps for general understanding, 4 fps for action recognition. Each frame is patchified identically to images. Temporal position embeddings encode frame index alongside 2D spatial position.

The unified sequence for a multimodal input is:

$$\mathcal{S} = [t_1, \dots, t_T] \oplus [v_1, \dots, v_P] \oplus [a_1, \dots, a_A] \oplus [\text{vid}_1, \dots, \text{vid}_F] \quad (1)$$

where $\oplus$ denotes concatenation with modality type embeddings prepended.

2.2 Cross-Modal Attention

The core innovation is a modified attention mechanism that allows unrestricted cross-modal attention within the same sequence. Rather than restricting vision tokens to attend only to vision tokens, Zen3-Omni allows each token to attend across the full multimodal context:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M_{\text{causal}}\right) V \quad (2)$$

where M_{causal} is a causal mask that prevents future text tokens from attending to past text tokens but permits bidirectional attention between vision, audio, and video tokens (which are treated as context, not generated sequence).

2.3 Temporal Sparse Attention for Video

Attending over 10-minute video at 1fps produces sequences of $600 \times 256 = 153,600$ vision tokens, making full attention intractable. Zen3-Omni uses a hierarchical temporal attention pattern:

- **Local window:** Each frame attends fully to adjacent ± 2 frames.
- **Strided global:** Every 10th frame attends to global frame summaries (mean-pooled representations of 10-frame segments).
- **Query-guided retrieval:** Text query tokens attend to all frame summary tokens, then retrieve top- k full-resolution frames for dense attention.

This reduces video attention from $O(F^2P^2)$ to $O(FP^2 + F^2/s)$ where $s = 10$ is the stride factor.

2.4 Zen MoDE Expert Routing with Modality Balance

The feed-forward layers use Zen MoDE expert routing with 64 experts, top-4 active per token. Third-generation routing adds a modality balance loss:

$$\mathcal{L}_{\text{balance}} = \alpha \sum_{e=1}^{64} \left(f_e - \frac{1}{64}\right)^2 + \beta \sum_{m \in \mathcal{M}} \text{KL}(p_m^e \| \bar{p}^e) \quad (3)$$

where f_e is the fraction of tokens routed to expert e , p_m^e is the routing distribution for modality m , and $\bar{p}^e$ is the average. This prevents individual experts from specializing exclusively on high-frequency text tokens, maintaining expert utilization across all modalities.

2.5 Streaming CTC Audio Decoder

For real-time audio, Zen3-Omni maintains a parallel CTC head that produces transcript hypotheses at 40ms intervals. The CTC output is fed back as text tokens into the main transformer context, enabling the model to reason over audio content while it is still being received.

3 Training

3.1 Pre-Training Data

Zen3-Omni pre-training uses a multimodal corpus of 3T text tokens and aligned multimodal data:

- Text: 1.8T tokens (same pipeline as Zen3-Small base model)
- Image-text pairs: 1.2B pairs from web alt-text, academic figures, and curated datasets
- Audio-text pairs: 120M samples (speech recognition, audio captioning, music description)
- Video-text pairs: 85M clips from instructional video, film, and synthetic data
- Interleaved documents: 400M multimodal documents with mixed modalities in context

3.2 Training Curriculum

Phase 1 — Modality encoders: Train image, audio, and video encoders independently with frozen LM backbone. Duration: 200B image tokens, 50B audio tokens, 20B video tokens.

Phase 2 — Unified pre-training: Jointly train all components on the full interleaved corpus with cross-modal attention enabled. Duration: 1.5T tokens equivalent.

Phase 3 — Instruction tuning: Fine-tune on 80M multimodal instruction samples covering all four modality combinations.

Phase 4 — RLHF: Preference optimization using a multimodal reward model trained on human ratings of response quality across all modality inputs.

3.3 Distillation from Zen3-VL

For vision-language tasks, Zen3-Omni uses Zen3-VL (72B) as a teacher, distilling visual understanding capabilities via logit matching on image-text pairs. This bridges the 10× parameter gap and brings Zen3-Omni VL performance within 5% of the 72B flagship.

4 Evaluation

4.1 Vision-Language

Table 1: Vision-language benchmarks

Model	Params	MMBench	MMMUS	TextVQA	OCRBench
Zen-Omni	7B	78.3	58.4	74.1	71.2
Zen2-Omni	7B	82.1	63.7	79.8	77.6
Zen3-Omni	7B	86.2	68.9	83.4	82.1

4.2 Audio Understanding

Table 2: Audio benchmarks

Model	AudioCaps	CLAP	LibriSpeech WER	MSVD-QA	AIR-Bench
Zen-Omni	0.481		3.8	61.3	71.4
Zen2-Omni	0.502		3.2	67.9	76.8
Zen3-Omni	0.523		2.7	73.2	81.9

4.3 Video Understanding

Table 3: Video understanding benchmarks

Model	VideoQA	EgoSchema	MVBench	Video-MME
Zen-Omni	68.4	52.1	71.2	61.3
Zen2-Omni	73.8	59.4	76.9	68.7
Zen3-Omni	79.4	66.2	83.1	74.8

4.4 Cross-Modal Reasoning

Table 4: Cross-modal reasoning: tasks requiring simultaneous use of 2+ modalities

Task	Zen3-Omni	Zen2-Omni
Audio-Visual QA (synchronized A+V)	74.3	63.1
Talking-Head Description (A+V+Text)	71.8	58.4
Chart with Narration Understanding	81.2	69.7
Live Lecture Comprehension	68.9	54.2

4.5 Streaming Latency

Table 5: End-to-end streaming inference latency (A100 80GB)

Input Mode	Time to First Token (ms)	Throughput (tok/s)
Text only	18	4200
Image + Text	62	3800
Audio stream (40ms)	43	3900
Video (1fps, 60s)	210	3100
All modalities	270	2800

5 Related Work

5.1 Multimodal Language Models

Early multimodal language models relied on frozen vision encoders with lightweight projection adapters connecting to a text backbone [1]. While effective for image captioning and VQA, these architectures treat vision tokens as second-class context, limiting deep cross-modal reasoning. Zen3-Omni’s unified token space builds on work showing that full cross-modal attention substantially improves tasks requiring joint reasoning.

5.2 Audio-Language Models

Audio integration with language models has advanced rapidly through mel spectrogram tokenization and learned audio codebooks. Streaming CTC integration, as used in Zen3-Omni, was pioneered for real-time speech translation systems and extended here to general audio understanding within a multimodal context.

5.3 Video Understanding

Long-form video understanding remains challenging due to the quadratic attention cost. Prior work addressed this through frame sampling, visual summarization, and hierarchical architectures. Zen3-Omni’s temporal sparse attention synthesizes these approaches into a unified mechanism compatible with the Zen MoDE backbone.

6 Conclusion

Zen3-Omni demonstrates that unified cross-modal reasoning in a 7B parameter model can achieve state-of-the-art results across vision, audio, video, and text simultaneously. The third-generation Zen MoDE architecture with modality-balanced expert routing, temporal sparse at-

tention, and streaming CTC decoding addresses the key challenges of prior generation multi-modal models.

Future work will extend to 3D scene understanding, tactile sensor inputs, and sub-100ms real-time video captioning for accessibility applications.

Acknowledgments

The authors thank Zoo Labs Foundation for multimodal dataset curation and compute allocation, and the Zen LM research community for benchmark contributions.

References

- [1] H. Liu et al., “Visual Instruction Tuning,” *NeurIPS*, 2023.
- [2] Zoo Labs Foundation, “ZIP-007: BitDelta — Quantization-Aware Delta Compression for Language Models,” *Zoo Improvement Proposals*, 2024.
- [3] Hanzo AI, “HIP-002: Active Semantic Optimization (ASO),” *Hanzo Improvement Proposals*, 2024.
- [4] Zoo Labs Foundation, “ZIP-001: Decentralized Semantic Optimization (DSO),” *Zoo Improvement Proposals*, 2024.