Zen-Omni: Efficient Multimodal AI with Identity Training

Zen LM Team Hanzo AI Zoo Labs Foundation

December 2024

Abstract

We present Zen-Omni, an efficient multimodal AI model built on the Qwen3-Omni-30B-A3B architecture with identity fine-tuning. Zen-Omni natively processes text, vision, and audio modalities through a unified Mixture-of-Experts (MoE) transformer architecture, activating only 3B of 30B total parameters per token. We apply LoRA-based identity training to establish consistent personality and capabilities, enabling deployment across consumer hardware while maintaining production-quality multimodal understanding. The model supports 119 text languages, 19 speech input languages, and 10 speech output languages, with applications in translation, dubbing, and cross-modal reasoning. Zen-Omni is released under Apache 2.0 license.

1 Introduction

Multimodal AI systems capable of understanding text, images, and audio simultaneously represent a significant advancement in artificial intelligence. However, deploying such systems efficiently while maintaining quality remains challenging.

Zen-Omni addresses this by leveraging the Qwen3-Omni architecture's Mixture-of-Experts design, which activates only a subset of parameters for each input, enabling efficient inference on consumer hardware. We further enhance the base model through identity fine-tuning, establishing the "Zen" persona with consistent behavior and specialized capabilities for translation and dubbing workflows.

Our contributions include:

- Identity-trained multimodal model with consistent persona
- Efficient deployment configurations (MLX 4-bit, GGUF quantized)
- Integration framework for video dubbing (zen-dub)
- Comprehensive training data for identity establishment

2 Architecture

Zen-Omni inherits the Thinker-Talker architecture from Qwen3-Omni:

2.1 Input Encoders

- **Audio Encoder**: 32 layers, 1280 dimensions, Whisper-style architecture
- **Vision Encoder**: 27 layers, 1152 dimensions, Vision Transformer
- **Text Embeddings**: 151,936 vocabulary size

2.2 Thinker (Multimodal LLM)

- 48 transformer layers with cross-modal attention
- 128 experts with 8 active per token (MoE)
- 30B total parameters, 3B active
- Extended thinking mode support (32K tokens)

2.3 Talker (Audio Generator)

- Streaming speech synthesis
- Code2Wav audio codec (16 quantizers, 2048 codebook)
- Real-time voice generation with prosody preservation

3 Training

3.1 Base Model

We start from `Qwen/Qwen3-Omni-30B-A3B-Instruct`, which provides multimodal instruction-following capabilities out of the box.

3.2 Identity Fine-Tuning

We apply LoRA fine-tuning with the following configuration:

- LoRA rank: 64, alpha: 128
- Target modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
- Training epochs: 3
- Batch size: 1 with gradient accumulation of 16
- Learning rate: 1e-4 with cosine scheduling

The identity dataset contains conversational examples establishing:

- Model name and creator attribution
- Capability descriptions (multimodal, translation, dubbing)
- Architecture knowledge (MoE, Thinker-Talker)
- Language support specifications

4 Evaluation

4.1 Multimodal Understanding

Zen-Omni maintains the base model's performance on standard benchmarks while adding consistent identity responses.

4.2 Deployment Efficiency

Format	Size	RAM	Tokens/sec
BF16	60GB	80GB+	5-10
MLX 8-bit	30GB	32GB	8-15
MLX 4-bit	15GB	20GB	10-20
GGUF Q4_K_M	15GB	20GB	10-20

Table 1: Performance on Apple Silicon (M2 Pro)

5 Applications

5.1 Omni Captioning

Audio and visual content captioning with cross-modal context understanding.

5.2 Speech Translation

Real-time translation across 119 languages with voice preservation for 10 output languages.

5.3 Video Dubbing

Integration with zen-dub enables lip-synchronized video translation using MuseTalk's VAE architecture.

5.4 Thinking Mode

Extended reasoning capabilities with up to 32K thinking tokens for complex problems.

6 Conclusion

Zen-Omni demonstrates that identity-trained multimodal models can be efficiently deployed on consumer hardware while maintaining consistent, production-quality behavior. The combination of MoE architecture, LoRA fine-tuning, and aggressive quantization enables sophisticated multimodal AI accessible to individual users and small teams.

6.1 Availability

Model weights, training code, and deployment configurations are available at `https://huggingface.co/zenlm/zen-omni` under Apache 2.0 license.