

Zen Multimodal Architecture: Unified Vision-Language-Audio

Technical Report v2025.03

Zen LM Research Team
research@zenlm.org

March 2025

Abstract

We present the Zen multimodal architecture, a unified framework for processing and generating text, images, and audio within a single model. Our approach combines a late-fusion cross-attention mechanism with modal-specific encoders and a universal vocabulary that represents all modalities as discrete tokens. Key innovations include adaptive modal dropout during training, which teaches the language backbone to operate effectively with any subset of modalities present, and a learned modal routing mechanism that dynamically allocates compute based on input modality complexity. The architecture achieves state-of-the-art results on MMBench (82.4), VideoQA (78.1%), AudioCaps retrieval (R@1 = 64.3%), and cross-modal retrieval (MSCOCO R@1 = 87.2%) while maintaining text-only performance within 0.3 points of the text-specialized baseline.

1 Introduction

Human cognition integrates perception across modalities seamlessly. Language models operating on text alone miss the rich structure available in images, video, and audio. Multimodal models must address three challenges:

1. **Alignment:** Mapping each modality into a shared semantic space
2. **Fusion:** Combining cross-modal information for joint reasoning
3. **Generation:** Producing outputs in any modality

Early multimodal models concatenated visual tokens to text context (ViLBERT, early GPT-4V style approaches). This is limited by context length and fails to capture fine-grained cross-modal attention patterns. Zen’s architecture uses late fusion: each modality is processed by a specialized encoder, and cross-modal attention is applied at the language backbone level.

2 Architecture Overview

2.1 Modal Encoders

2.1.1 Vision Encoder

The vision encoder is a ViT-L/14 variant with 307M parameters, trained from scratch on 2B image-text pairs. It produces patch embeddings of dimension 1024 at 14×14 pixel patches. For a 224×224 image, this yields $16 \times 16 = 256$ visual tokens.

Dynamic resolution processing supports arbitrary input sizes:

$$N_{\text{patches}} = \lceil H/14 \rceil \times \lceil W/14 \rceil \tag{1}$$

with 2D rotary positional embeddings that interpolate smoothly to unseen resolutions.

2.1.2 Audio Encoder

The audio encoder converts raw waveforms to mel spectrograms and processes them with a 1D convolutional front-end followed by a transformer:

$$\text{mel}[t, f] = \log \left(\sum_k w[k] \cdot |X[t, k]|^2 + \epsilon \right) \quad (2)$$

with 80 mel bins, 10ms frame shift, and 25ms window. The transformer produces audio tokens at 50Hz (one token per 20ms). Audio sequences up to 30 seconds are supported natively (1500 tokens).

2.1.3 Universal Tokenizer

Text tokens follow the standard BPE vocabulary (150K tokens). Visual and audio tokens are produced by vector-quantized encoders with codebook size 8192:

$$q(z) = \arg \min_{e \in \mathcal{E}} \|z - e\|_2 \quad (3)$$

where $\mathcal{E}$ is the learned codebook. Visual codebooks are trained on image reconstruction; audio codebooks on waveform reconstruction with a GAN loss. This gives a unified discrete vocabulary of $150\text{K} + 2 \times 8192 = 166\text{K}$ tokens across all modalities.

2.2 Cross-Modal Attention Fusion

The language backbone (Zen-7B) processes the combined sequence of text, visual, and audio tokens. To prevent text processing from being dominated by the larger visual/audio token sequences, we use gated cross-modal attention:

$$\text{CrossAttn}(Q_{\text{text}}, K_{\text{vis}}, V_{\text{vis}}) = \text{softmax} \left(\frac{Q_{\text{text}} K_{\text{vis}}^\top}{\sqrt{d}} \right) V_{\text{vis}} \quad (4)$$

A gating scalar α interpolates between self-attention output and cross-attention:

$$h = (1 - \alpha) \cdot \text{SelfAttn}(h) + \alpha \cdot \text{CrossAttn}(h, V_{\text{vis}}) \quad (5)$$

with α learned per layer via a sigmoid gate initialized to 0. This allows the model to gradually learn cross-modal integration rather than immediately disrupting pretrained text representations.

2.3 Modal Dropout

During training, we randomly drop input modalities with probability:

- Drop vision: 15%
- Drop audio: 20%
- Drop both: 5%
- Keep all: 60%

Modal dropout teaches the model to extract maximum information from available modalities and prevents over-reliance on any single modality. It also enables zero-shot transfer to text-only, vision-only, and audio-only inference modes.

3 Training

3.1 Stage 1: Modal Encoder Pretraining

Vision encoder: trained on 2B image-text pairs (LAION-5B subset) with contrastive CLIP-style loss for 100K steps.

Audio encoder: trained on 500K audio-text pairs (AudioCaps, AudioSet with captions) with contrastive loss for 50K steps.

3.2 Stage 2: Unified Backbone Training

The pretrained modal encoders are connected to the Zen-7B language backbone. We freeze encoder weights and train only the cross-attention gates and projection layers for 50K steps on multimodal instruction data:

Data Type	Samples	Mix Weight
Image + text Q&A	1.2M	40%
Image captioning	800K	20%
Video + text	400K	15%
Audio + text	300K	10%
Text only	2.0M	15%

Table 1: Stage 2 multimodal training data composition.

3.3 Stage 3: End-to-End Fine-Tuning

All components are fine-tuned jointly with a lower learning rate ($\eta = 10^{-5}$) for 20K steps. The text-only loss coefficient is upweighted by $2\times$ to prevent catastrophic forgetting of text capabilities.

4 Experiments

4.1 Visual Understanding: MMBench

Model	Overall	LR	AR	RR	FP
Zen-7B-Multimodal	82.4	80.1	84.3	79.8	85.2
Zen-32B-Multimodal	85.1	83.4	86.8	82.3	88.1

Table 2: MMBench results. LR=Logical Reasoning, AR=Attribute Recognition, RR=Relation Reasoning, FP=Fine-grained Perception.

4.2 Video Understanding: VideoQA

Model	MSVD-QA	MSRVTT-QA	ActivityNet-QA	Avg.
Zen-7B-Multimodal	79.3%	76.8%	78.2%	78.1%
Zen-32B-Multimodal	82.1%	79.4%	81.0%	80.8%

Table 3: VideoQA accuracy across benchmarks.

4.3 Audio Understanding: AudioCaps

Model	R@1	R@5	R@10	CIDEr
Zen-7B-Multimodal	64.3	88.2	94.1	82.4

Table 4: AudioCaps text-to-audio retrieval and captioning results.

4.4 Cross-Modal Retrieval: MSCOCO

Model	Image→Text			Text→Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Zen-7B-Multimodal	87.2	97.8	99.1	72.4	91.3	96.2

Table 5: MSCOCO 5K test set cross-modal retrieval.

4.5 Text-Only Performance Preservation

Model	MMLU	HumanEval	MATH	MT-Bench
Zen-7B (text only)	85.3	78.2	67.4	8.62
Zen-7B-Multimodal	85.1	78.0	67.2	8.59
Δ	-0.2	-0.2	-0.2	-0.03

Table 6: Text performance of multimodal vs. text-only model. Modal dropout prevents degradation.

5 Analysis

5.1 Ablation: Cross-Attention Fusion Strategy

Fusion Strategy	MMBench	Text MMLU
Prefix concatenation	76.2	83.1
Early fusion (all layers)	79.8	82.4
Late fusion (top 1/3 layers)	81.3	84.7
Gated cross-attention (ours)	82.4	85.1

Table 7: Fusion strategy ablation. Gated cross-attention best preserves text while gaining vision.

5.2 Effect of Modal Dropout Rate

Dropout Rate	MMBench	Text MMLU	Text-Only Inference
0%	83.1	83.4	79.2
10%	82.8	84.6	83.1
20% (default)	82.4	85.1	84.9
40%	80.9	85.2	85.3

Table 8: Effect of modal dropout on multimodal and text-only performance.

Higher dropout rates improve text-only inference robustness at a small cost to peak multimodal performance. The 20% default balances both objectives.

6 Conclusion

The Zen multimodal architecture demonstrates that a unified vision-language-audio model can achieve strong performance across all modalities with minimal sacrifice to text-only capability. Gated cross-attention fusion and modal dropout are the key methodological contributions enabling this balance. The architecture naturally extends to video (sequential image frames) and future modalities via the universal discrete tokenizer.

References

- [1] Radford et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ICML*.
- [2] Dosovitskiy et al. (2021). An Image is Worth 16x16 Words. *ICLR*.
- [3] Gemini Team (2024). Gemini: A Family of Highly Capable Multimodal Models. *arXiv:2312.11805*.
- [4] Liu et al. (2023). Visual Instruction Tuning. *NeurIPS*.
- [5] Li et al. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *ICML*.