

Zen AI Model Family

Zen-World

Packaging Wan2.1 Text-to-Video Generation for the Zen Stack

Technical Report

Hanzo AI Research Team*

Zoo Labs Foundation†

June 2026

Abstract

Zen-World is a packaging and integration of **Wan2.1-T2V-14B**, the open-source text-to-video diffusion model developed and released by Alibaba’s Team Wan, into the Zen model stack. Zen-World is *not* a from-scratch model: it redistributes the upstream Wan-AI/Wan2.1-T2V-14B weights and architecture under the permissive Apache 2.0 license, adding Zen-stack loaders, configuration, and a unified inference interface. The underlying model is a ~ 14 -billion-parameter Diffusion Transformer (DiT) trained with a flow-matching objective, paired with the Wan2.1 3D causal variational autoencoder (Wan-VAE) for spatio-temporal latent compression and the multilingual umT5 text encoder for prompt conditioning. It generates 480P and 720P video from English or Chinese text prompts. This report describes the real architecture as released upstream, attributes it to its authors, and documents how it is integrated into Zen. All architectural and capability claims here are those of the upstream Wan2.1 release; we do not report any independent benchmark numbers of our own.

Attribution and License

Upstream model. Zen-World redistributes Wan-AI/Wan2.1-T2V-14B, part of the *Wan* series of open large-scale video generative models from Alibaba’s Team Wan [1].

License. The Wan2.1 models and source code are released under the **Apache License 2.0**. Zen-World preserves this license and the upstream copyright and attribution notices. No relicensing is claimed; redistribution is permitted because Apache 2.0 is a permissive license.

What Zen adds. Zen-World contributes *packaging only*: Zen-stack model loaders, a consistent configuration and weights layout, and a unified text-to-video inference API consistent with the rest of the Zen family. It does not modify the model architecture or retrain the weights.

Contents

Attribution and License	1
1 Introduction	2
1.1 Model Overview	2
2 Architecture	2
2.1 Wan-VAE: 3D Causal Video Autoencoder	2
2.2 umT5 Text Encoder	3
2.3 Diffusion Transformer (DiT)	3
2.4 Flow-Matching Objective	3

*research@hanzo.ai

†foundation@zoo.ngo

3	Capabilities (as reported upstream)	3
4	Integration into the Zen Stack	4
5	Related Work	4
6	Conclusion	5

1 Introduction

Text-to-video generation synthesizes a short video clip from a natural-language description. Modern systems follow the diffusion-transformer paradigm: a transformer denoiser operates in a compressed video latent space produced by a video autoencoder, conditioned on text embeddings from a language encoder, and is trained to reverse a noising (or flow-matching) process [1].

Zen-World packages one such system—**Wan2.1-T2V-14B**—into the Zen stack. Wan2.1 was released by Alibaba’s Team Wan as an open, Apache-2.0-licensed family of video generative models [1]. The 14B text-to-video variant is the model redistributed here. Zen-World provides Zen-native loading and a unified inference interface; the generative capability, weights, and architecture are entirely those of the upstream release.

This report (1) attributes the model to its authors and states its license, (2) describes the real Wan2.1 architecture and its components, (3) clarifies the scope of Zen’s integration, and (4) reports only capabilities and figures stated by the upstream release, with attribution.

1.1 Model Overview

Property	Value
Upstream model	Wan-AI/Wan2.1-T2V-14B (Alibaba Team Wan)
Task	Text-to-video generation
License	Apache 2.0
Architecture	Diffusion Transformer (DiT), flow matching
Denoiser parameters	~14B
Latent autoencoder	Wan2.1 3D causal VAE (Wan-VAE)
Text encoder	umT5 (multilingual; English / Chinese prompts)
Output resolutions	480P (832×480), 720P (1280×720)
Zen contribution	Packaging / integration into the Zen stack

Table 1: Zen-World summary. Architecture and capability values are those of upstream Wan2.1.

2 Architecture

Wan2.1 follows the mainstream diffusion-transformer design and is composed of three primary components [1]: a 3D causal VAE (*Wan-VAE*), a multilingual umT5 text encoder, and a flow-matching Diffusion Transformer (DiT) denoiser. Figure-level overview:

1. **Wan-VAE** encodes video into a compact spatio-temporal latent and decodes generated latents back to pixels.
2. **umT5** encodes the text prompt into embeddings.
3. **DiT** denoises the video latent, conditioned on the text embeddings via cross-attention, under a flow-matching objective.

2.1 Wan-VAE: 3D Causal Video Autoencoder

Generating video directly in pixel space is prohibitively expensive, so Wan2.1 operates in a compressed latent space produced by **Wan-VAE**, a 3D *causal* variational autoencoder designed for video [1]. The VAE performs spatio-temporal compression with a temporal downsampling factor of $t = 4$ and spatial downsampling factors of $h = w = 8$, while preserving temporal causality so that motion remains smooth across frames. The latent has 16 channels (the DiT’s input and output channel dimension; see Table 2). The causal design lets the VAE encode and decode long videos without breaking temporal consistency.

2.2 umT5 Text Encoder

Prompts are encoded by **umT5**, a multilingual T5-family text encoder. The resulting text embeddings condition the DiT through cross-attention. The multilingual encoder is one reason Wan2.1 accepts both English and Chinese prompts, and—per the upstream release—Wan2.1 is described as the first video model able to generate both Chinese and English text within the rendered video [1]. The text sequence length used by the 14B configuration is 512 tokens (Table 2).

2.3 Diffusion Transformer (DiT)

The denoiser is a Diffusion Transformer. It processes the noised video latent as a sequence of patch tokens through a stack of transformer blocks. Each block (the *WanAttentionBlock*) combines, per the upstream design, vision self-attention over the video tokens, text-to-vision cross-attention that injects the umT5 prompt embeddings, and a feed-forward layer, with timestep conditioning applied to modulate the block [1].

The exact configuration of the redistributed 14B text-to-video model, taken directly from the upstream model configuration, is given in Table 2.

Hyperparameter	Value (T2V-14B)
Model type	t2v (text-to-video)
Model / embedding dimension (dim)	5120
Transformer layers (num_layers)	40
Attention heads (num_heads)	40
Feed-forward dimension (ffn_dim)	13824
Latent input / output channels (in_dim / out_dim)	16
Frequency embedding dimension (freq_dim)	256
Text sequence length (text_len)	512
LayerNorm epsilon (eps)	1×10^{-6}

Table 2: Wan2.1-T2V-14B DiT configuration, from the upstream `config.json`.

2.4 Flow-Matching Objective

Rather than a traditional discrete-noise diffusion schedule, Wan2.1 is trained with **flow matching** [2, 1]. Flow matching defines intermediate states by linear interpolation between a sample and noise and trains the network to predict the corresponding velocity (the time derivative of the interpolation path). Concretely, for a clean latent x_1 and noise $x_0 \sim \mathcal{N}(0, I)$, an intermediate latent is

$$x_t = (1 - t)x_0 + tx_1, \quad t \in [0, 1], \quad (1)$$

and the model v_θ is trained to regress the target velocity $x_1 - x_0$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, x_0, x_1, c} \|v_\theta(x_t, t, c) - (x_1 - x_0)\|_2^2, \quad (2)$$

where c denotes the umT5 text conditioning. At inference, video latents are generated by integrating the learned velocity field from noise, then decoded to pixels by Wan-VAE.

3 Capabilities (as reported upstream)

The capabilities below are those stated by the upstream Wan2.1 release [1]; they are reproduced here with attribution. Zen-World does not add or independently re-measure these figures.

- **Output.** Generates short video clips at 480P (832×480) and 720P (1280×720) from text prompts.
- **Multilingual prompts and in-video text.** Accepts English and Chinese prompts; described upstream as the first video model able to render both Chinese and English text in the generated video.
- **Model family.** The Wan2.1 release also includes a smaller T2V-1.3B model and image-to-video (I2V), first-last-frame (FLF2V), and video-editing (VACE) variants; Zen-World redistributes the T2V-14B variant.
- **Efficiency (sibling model).** The upstream release reports that the smaller T2V-1.3B model can generate a 5-second 480P clip using about 8.19 GB of VRAM, enabling consumer-GPU use; the 14B model redistributed here is correspondingly heavier.
- **Evaluation.** The upstream report states the 14B model was evaluated on an internal suite of 1,035 prompts across 14 dimensions and that it is competitive with leading open-source and commercial systems. We reproduce this claim as reported and do not restate specific scores here.

4 Integration into the Zen Stack

Zen-World’s contribution is integration, not modeling. Specifically, Zen-World provides:

- Zen-stack loading of the upstream Wan2.1-T2V-14B weights (DiT, Wan-VAE, umT5).
- A consistent on-disk configuration and weights layout aligned with the Zen family.
- A unified text-to-video inference interface consistent with other Zen models.

No architectural changes are made and the weights are not retrained. The example below illustrates the intended Zen-stack interface for text-to-video generation.

```

1 from zen import ZenWorld
2
3 # Loads the redistributed Wan2.1-T2V-14B weights (DiT + Wan-VAE + umT5)
4
5 model = ZenWorld.from_pretrained("zenlm/zen-world")
6
7 # Text-to-video generation.
8 video = model.generate(
9     prompt="A red panda walking through a snowy bamboo forest, cinematic lighting",
10     resolution="720p",          # 1280x720; also "480p" (832x480)
11     num_frames=81,
12     fps=16,
13 )
14 video.save("output.mp4")

```

Listing 1: Zen-World text-to-video (packaging of Wan2.1-T2V-14B)

5 Related Work

Zen-World is a redistribution of Wan2.1 [1], which belongs to the broader line of diffusion-transformer video generators. The diffusion-transformer (DiT) backbone derives from the scalable transformer denoiser of Peebles and Xie [3]. Flow matching, the training objective Wan2.1

uses in place of a discrete noise schedule, was introduced by Lipman et al. [2]. The Wan technical report [1] details the Wan-VAE 3D causal autoencoder, the umT5 conditioning, and the model family.

6 Conclusion

Zen-World is the Zen stack’s packaging of **Wan2.1-T2V-14B**, the Apache-2.0 text-to-video diffusion model released by Alibaba’s Team Wan. The real system is a $\sim 14\text{B}$ Diffusion Transformer trained with flow matching, operating in the latent space of the Wan2.1 3D causal VAE and conditioned by the multilingual umT5 text encoder, producing 480P and 720P video from English or Chinese prompts. Zen’s contribution is integration: loaders, configuration, and a unified inference interface. All architecture and capability statements in this report are attributed to the upstream Wan2.1 release, and no independent benchmark numbers are claimed.

References

- [1] Team Wan (Alibaba), “Wan: Open and Advanced Large-Scale Video Generative Models,” arXiv:2503.20314, 2025. Models and code: <https://github.com/Wan-Video/Wan2.1>; weights: <https://huggingface.co/Wan-AI/Wan2.1-T2V-14B> (Apache 2.0).
- [2] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow Matching for Generative Modeling,” ICLR, 2023. arXiv:2210.02747.
- [3] W. Peebles and S. Xie, “Scalable Diffusion Models with Transformers,” ICCV, 2023. arXiv:2212.09748.