

Zen AI Model Family

Zen-Director

Single-GPU Unified Text- and Image-to-Video Generation

A Packaging of Alibaba’s Wan2.2-TI2V-5B (Apache-2.0)

Technical Report

Hanzo AI Research Team*

Zoo Labs Foundation†

2026

Abstract

Zen-Director is a permissively licensed redistribution and integration of **Wan2.2-TI2V-5B**, the open-source unified text- and image-to-video generation model released by the Alibaba Wan team (Wan-AI/Wan2.2-TI2V-5B) under the **Apache-2.0 license** [1, 2]. Zen-Director is *not* a from-scratch model: it packages Wan2.2-TI2V-5B—a 5-billion-parameter *dense* diffusion transformer trained with flow matching, paired with the high-compression Wan2.2 3D causal VAE—into the Zen model family with consistent tooling, weights distribution, and serving integration. The defining property of Wan2.2-TI2V-5B is that it generates 720P video at 24 fps on a *single consumer-grade GPU* (e.g. an NVIDIA RTX 4090 with under 24 GB of VRAM), generating a 5-second 720P clip in under nine minutes without specialized optimization [1, 2]. A single unified framework natively serves both text-to-video (T2V) and image-to-video (I2V). This report documents the model we redistribute, its true architecture, its capabilities, and how it is integrated into the Zen stack. All architectural facts and figures below are reported by the upstream Wan team; we add no independent benchmarks.

Contents

1	Introduction	3
1.1	Provenance and License	3
1.2	Model Overview	3
1.3	Key Capabilities	4
2	Architecture	4
2.1	5B Dense Diffusion Transformer	4
2.2	High-Compression Wan2.2 3D Causal VAE	4
2.3	umT5 Text Encoder	4
2.4	Flow-Matching Training Objective	4
2.5	Unified T2V + I2V Framework	5
3	Inference and Integration	5
3.1	Running Zen-Director	5
3.2	Hardware Footprint	5
3.3	Role in the Zen Stack	5
4	Reported Results	6

*research@hanzo.ai

†foundation@zoo.ngo

5	Licensing and Attribution	6
6	Related Work	6
7	Conclusion	6

1 Introduction

Open-source video generation has reached the point where high-resolution, high-frame-rate synthesis is feasible on a single consumer GPU. The Alibaba Wan team’s **Wan2.2** family [2, 3] is a leading example of this shift. Within that family, **Wan2.2-TI2V-5B** is a compact 5-billion-parameter *dense* model that unifies text-to-video and image-to-video generation and runs within the memory budget of widely available hardware. Crucially, both the code and the weights are released under the **Apache-2.0 license** [1, 2], which permits commercial redistribution and modification with attribution.

Zen-Director is our packaging of Wan2.2-TI2V-5B. We make no claim to having designed, trained, or fine-tuned a new architecture. Our contribution is integration: distributing the upstream Apache-2.0 weights through the Zen model family, wiring the model into the Zen serving and tooling stack, and providing a consistent interface alongside other Zen models. This document therefore describes the *real* Wan2.2-TI2V-5B architecture and capabilities so that users of Zen-Director understand exactly what they are running and under what license.

1.1 Provenance and License

Property	Value
Upstream model	Wan2.2-TI2V-5B
Upstream author	Alibaba Wan team (Wan-Video / Wan-AI)
Upstream repository	Wan-AI/Wan2.2-TI2V-5B (Hugging Face)
License	Apache-2.0
Zen-Director role	Packaging / integration, not from-scratch training
Modality	Unified text-to-video (T2V) + image-to-video (I2V)

Table 1: Zen-Director provenance. Zen-Director redistributes the upstream model under its original Apache-2.0 license with attribution to the Alibaba Wan team.

1.2 Model Overview

Property	Value (as reported upstream [1, 2])
Parameters	5B (dense; not Mixture-of-Experts)
Architecture	Diffusion Transformer (DiT) + Wan2.2 3D causal VAE
Text encoder	umT5 (multilingual)
Training objective	Flow matching
VAE compression ($T \times H \times W$)	$4 \times 16 \times 16$ (overall rate ≈ 64)
With patchification	$4 \times 32 \times 32$
Output resolution	720P (1280×704 or 704×1280)
Frame rate	24 fps
Clip length	up to 5 seconds
Text input	up to 512 tokens
Image input (I2V)	up to 1280×704
GPU requirement	single consumer GPU, < 24 GB VRAM (e.g. RTX 4090)
Reported speed	5-second 720P clip in < 9 minutes (single GPU, no special optimization)

Table 2: Wan2.2-TI2V-5B specifications as reported by the upstream Wan team. Zen-Director redistributes these weights unchanged.

1.3 Key Capabilities

- **Unified text-to-video:** Generate a 720P, 24 fps clip (up to 5 seconds) from a text prompt of up to 512 tokens.
- **Unified image-to-video:** Animate a reference image (up to 1280×704) into a coherent video within the same single framework.
- **Single-GPU operation:** Run end-to-end on one consumer-grade GPU with under 24 GB of VRAM, enabled by the high-compression 3D causal VAE.
- **Permissive licensing:** Apache-2.0 weights and code, suitable for commercial redistribution with attribution.

2 Architecture

Zen-Director runs the Wan2.2-TI2V-5B model unchanged. Its architecture follows the Wan design lineage [3, 2]: a 3D causal variational autoencoder (VAE), a umT5 text encoder, and a diffusion transformer (DiT), trained with flow matching. We describe each component as documented upstream.

2.1 5B Dense Diffusion Transformer

The generative backbone is a *dense* diffusion transformer with approximately 5 billion parameters [1, 2]. Unlike the larger Wan2.2 A14B variants—which use a Mixture-of-Experts design with a high-noise expert for early denoising and a low-noise expert for later refinement—the TI2V-5B model is a single dense network with no expert routing. The DiT operates on the compressed video latents produced by the VAE: a patchifying module forms tokens from the latent, a stack of transformer blocks denoises them while attending to the text embedding via cross-attention, and an unpatchifying module reconstructs the latent.

2.2 High-Compression Wan2.2 3D Causal VAE

The capability that lets a 720P 24 fps video model fit on a single 24 GB GPU is the high-compression **Wan2.2-VAE**, a 3D causal autoencoder [2, 1]. It achieves a spatio-temporal ($T \times H \times W$) compression ratio of $4 \times 16 \times 16$, raising the overall information compression rate to roughly 64 while maintaining high-quality reconstruction. Combined with the patchification applied inside the DiT, the effective spatial reduction reaches $4 \times 32 \times 32$. The 3D *causal* structure preserves temporal causality across frames and substantially reduces the memory footprint of the latent sequence, which is what makes single-GPU 720P generation practical.

2.3 umT5 Text Encoder

Text conditioning uses the umT5 multilingual text encoder [3], which encodes the input prompt (up to 512 tokens) into embeddings consumed by the DiT through cross-attention. The Wan team selected umT5 for its strong multilingual (notably Chinese and English) encoding and favorable training convergence.

2.4 Flow-Matching Training Objective

The diffusion transformer is trained with **flow matching** [3]. Given a target latent x_1 , a noise sample x_0 , and a timestep $t \in [0, 1]$, the training input is the linear interpolation

$$x_t = (1 - t)x_0 + tx_1, \tag{1}$$

and the model is trained to predict the velocity field

$$v_t = x_1 - x_0. \quad (2)$$

At inference, samples are produced by integrating the learned velocity field, avoiding iterative per-step velocity re-prediction schemes used by some alternative diffusion formulations.

2.5 Unified T2V + I2V Framework

A single model serves both modalities [1, 2]. For text-to-video the model is conditioned on the umT5 text embedding alone; for image-to-video the reference image (up to 1280×704) provides additional conditioning. Both paths share the same VAE, DiT, and flow-matching sampler, so no separate model or checkpoint is required to switch between text-driven and image-driven generation.

3 Inference and Integration

3.1 Running Zen-Director

Because Zen-Director is the upstream Wan2.2-TI2V-5B model, it is invoked through the standard Wan2.2 generation interface. The example below illustrates the conceptual call (refer to the upstream repository for the authoritative API and flags) [2].

```
1 # Text-to-video: 720P, 24 fps, up to 5 seconds, single 24GB GPU
2 generate(
3     task="ti2v-5B",
4     size="1280*704",
5     prompt="a_lighthouse_at_dusk, waves_breaking, slow_push-in",
6 )
7
8 # Image-to-video: animate a reference image in the same framework
9 generate(
10    task="ti2v-5B",
11    size="1280*704",
12    image="reference.png",
13    prompt="gentle_camera_drift, wind_moving_through_grass",
14 )
```

Listing 1: Invoking Zen-Director (Wan2.2-TI2V-5B)

3.2 Hardware Footprint

The upstream Wan team reports that Wan2.2-TI2V-5B generates a 5-second 720P (24 fps) clip in under nine minutes on a single consumer-grade GPU, without specialized optimization, within a sub-24 GB VRAM budget such as an NVIDIA RTX 4090 [1, 2]. This single-GPU profile is the primary reason Zen-Director is suitable for self-hosted and on-premise deployments where multi-GPU clusters are unavailable.

3.3 Role in the Zen Stack

Within the Zen model family, Zen-Director is the video-generation component. It consumes a text prompt (and optionally a reference image) and produces video, complementing other Zen models that handle language and image tasks. Integration consists of distributing the Apache-2.0 weights, exposing the unified T2V/I2V interface through Zen tooling, and providing consistent serving alongside the rest of the family. No retraining or architectural modification is performed.

4 Reported Results

The Wan team positions Wan2.2-TI2V-5B as one of the fastest 720P@24fps open models currently available and reports favorable quality on its internal Wan-Bench 2.0 evaluation relative to contemporary models [2]. The public model card does not publish a specific per-metric VBench table for the TI2V-5B variant, and *we deliberately report no independent benchmark numbers of our own*: Zen-Director redistributes the upstream weights without modification, so any quantitative claims should be taken from, and attributed to, the upstream Wan releases [1, 2, 3].

5 Licensing and Attribution

Zen-Director is governed by the **Apache-2.0 license** of the upstream Wan2.2-TI2V-5B release [1, 2]. Redistribution under the Zen name retains this license and includes attribution to the Alibaba Wan team as the originating authors of the model, weights, and architecture. Apache-2.0 permits commercial use, modification, and redistribution provided the license and notices are preserved. Users remain responsible for ensuring that generated content complies with applicable laws and does not infringe third-party rights.

6 Related Work

Wan2.2-TI2V-5B belongs to the broader Wan family of open large-scale video generative models [3, 2], which also includes the larger A14B Mixture-of-Experts text-to-video and image-to-video models. The Wan line builds on the diffusion-transformer approach to video synthesis, combining a 3D causal VAE for efficient latent representation, a multilingual text encoder, and flow-matching training. Zen-Director’s specific value is in making the compact, single-GPU TI2V-5B configuration of this family conveniently available and integrated within the Zen stack.

7 Conclusion

Zen-Director is an honest redistribution of Alibaba’s Apache-2.0 Wan2.2-TI2V-5B model: a 5B dense diffusion transformer with a high-compression 3D causal VAE that generates 720P, 24 fps video from text or images on a single consumer GPU. It is not a from-scratch architecture and introduces no new benchmarks; its contribution is packaging and integration. By attributing the upstream Wan team and preserving the Apache-2.0 license, Zen-Director brings state-of-the-art, permissively licensed, single-GPU video generation into the Zen model family with full transparency about its provenance.

References

- [1] Alibaba Wan team, “Wan2.2-TI2V-5B,” Hugging Face model card, Wan-AI/Wan2.2-TI2V-5B. <https://huggingface.co/Wan-AI/Wan2.2-TI2V-5B>. License: Apache-2.0.
- [2] Wan-Video, “Wan2.2: Open and Advanced Large-Scale Video Generative Models,” GitHub repository. <https://github.com/Wan-Video/Wan2.2>. License: Apache-2.0.
- [3] Wan team, “Wan: Open and Advanced Large-Scale Video Generative Models,” arXiv:2503.20314, 2025. <https://arxiv.org/abs/2503.20314>.