

Zen-Video-I2V: Packaging Wan2.2-I2V-A14B for Image-to-Video Generation in the Zen Stack

Technical Whitepaper

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

2026

Abstract

Zen-Video-I2V is a redistribution and integration of **Wan2.2-I2V-A14B**, the open-source image-to-video (I2V) generation model released by the Alibaba Wan team (**Wan-AI/Wan2.2-I2V-A14B**) under the **Apache-2.0** license. It is *not* a from-scratch model: the weights, architecture, and training are the work of the Wan team, and Zen-Video-I2V repackages them for use within the Zen stack (consistent tooling, inference defaults, provenance, and safety integration). This whitepaper accurately describes the underlying Wan2.2 architecture — a latent diffusion transformer (DiT) operating over a 3D causal VAE, trained with a flow-matching objective, and using an A14B mixture-of-experts (MoE) design that splits denoising between a high-noise and a low-noise expert (approximately 27B total parameters, roughly 14B active per step). We summarize the model’s reported capabilities as published by the Wan team, document our packaging and integration choices, and discuss licensing, provenance, and safety. We deliberately do not report independent benchmark numbers; where quantitative claims appear, they are attributed to the upstream Wan team.

Contents

1	Introduction and Attribution	3
1.1	What Zen-Video-I2V is	3
1.2	The image-to-video task	3
1.3	Model overview	3
2	Architecture (Wan2.2-I2V-A14B)	3
2.1	Latent diffusion with a 3D causal VAE	4
2.2	Diffusion transformer (DiT) backbone	4
2.3	Flow-matching training objective	4
2.4	A14B mixture-of-experts: high-noise and low-noise experts	4
2.5	Image conditioning	5
3	Capabilities and Reported Results	5
3.1	Outputs	5
3.2	Training-data scale (upstream-reported)	5
3.3	Benchmarks (upstream-reported, attributed)	5

4	Zen Integration	5
4.1	Packaging	5
4.2	Inference characteristics	6
4.3	API interface	6
5	Licensing	6
6	Provenance and Safety	6
6.1	Content provenance	6
6.2	Identity and consent	7
6.3	Misuse mitigation	7
7	Related Work	7
8	Conclusion	7

1 Introduction and Attribution

1.1 What Zen-Video-I2V is

Zen-Video-I2V is a packaging of the upstream open-source model **Wan2.2-I2V-A14B** by Alibaba’s Wan team. The original model, weights, training, and architecture are entirely the work of the Wan team and are distributed at **Wan-AI/Wan2.2-I2V-A14B** on Hugging Face under the **Apache-2.0** license. Zen-Video-I2V exists to integrate this model into the Zen stack: it provides consistent command-line and API surfaces, sensible inference defaults, content-provenance (C2PA) emission, and safety wiring, while leaving the model architecture and weights unmodified except where explicitly noted.

This document supersedes earlier drafts that incorrectly described a from-scratch “Zen MoDE” architecture and reported fabricated benchmark results. No such architecture exists, and those numbers were not real. The remainder of this paper describes the actual Wan2.2-I2V-A14B model and our integration of it.

1.2 The image-to-video task

Image generation learns the distribution of plausible spatial configurations. Video generation must additionally learn plausible temporal dynamics: how objects move, deform, and interact over time while maintaining spatial and temporal coherence across frames. Image-to-video (I2V) constrains the problem by conditioning on a real input image (a photograph, a generated image, or concept art), which fixes the starting frame and shifts the task to learning temporal evolutions consistent with that start state. Wan2.2-I2V-A14B targets this setting, optionally also accepting a text prompt to steer the motion and content of the generated clip.

1.3 Model overview

Table 1: Zen-Video-I2V / Wan2.2-I2V-A14B specification, as reported by the Wan team. Quantitative items are upstream-reported, not independently measured by Zen.

Item	Value
Upstream model	Wan2.2-I2V-A14B (Alibaba Wan)
Upstream repository	Wan-AI/Wan2.2-I2V-A14B
License	Apache-2.0
Task	Image-to-video (image, optionally text-conditioned)
Architecture	Latent diffusion transformer (DiT) + 3D causal VAE
Training objective	Flow matching
Expert design	A14B mixture-of-experts (high-noise + low-noise)
Total parameters	$\approx 27\text{B}$ (two $\approx 14\text{B}$ experts)
Active parameters / step	$\approx 14\text{B}$
VAE compression ($T \times H \times W$)	$4 \times 8 \times 8$ (Wan VAE)
Output resolutions	480P (832×480) and 720P (1280×720)
Default clip	81 frames at 16 fps (≈ 5 s)
Text encoder	umT5-XXL

2 Architecture (Wan2.2-I2V-A14B)

The architecture described in this section is that of the upstream Wan2.2 model; it is reproduced here for completeness and is attributed to the Wan team [1, 2].

2.1 Latent diffusion with a 3D causal VAE

Wan2.2 follows the latent-diffusion paradigm: a 3D causal variational autoencoder (the Wan VAE) compresses video into a compact spatio-temporal latent representation, and a diffusion transformer models the distribution over those latents. For the A14B models, the VAE applies a $T \times H \times W$ downsampling factor of $4 \times 8 \times 8$ — $4 \times$ along time and $8 \times$ along each spatial axis. (The separate, smaller TI2V-5B variant of Wan2.2 instead uses a higher-compression Wan2.2-VAE with a $4 \times 16 \times 16$ ratio, reaching $4 \times 32 \times 32$ after patchification; this whitepaper concerns the A14B I2V model, not TI2V-5B.)

Operating in latent space rather than pixel space is what makes high-resolution, multi-second video generation computationally tractable: the diffusion transformer denoises a much smaller latent volume, and the VAE decoder reconstructs pixels at the end.

2.2 Diffusion transformer (DiT) backbone

The denoiser is a diffusion transformer (DiT) [3] operating over the VAE latent tokens. As is standard for DiT-based video models, the backbone uses self-attention over the flattened spatio-temporal latent tokens together with cross-attention to the conditioning signal, with timestep conditioning applied throughout. Text conditioning, when a prompt is supplied, is provided through a umT5-XXL text encoder [4]; the Wan codebase and model card credit umT5-XXL as the text encoder used in the Wan family. Image conditioning supplies the fixed first-frame content that the I2V model animates.

2.3 Flow-matching training objective

Wan2.2 is trained with a flow-matching objective rather than the classical ϵ -prediction DDPM loss. Flow matching learns a continuous-time velocity field that transports samples between the noise distribution and the data distribution; the model is trained to predict the velocity along the probability path, and generation integrates this learned velocity field at inference time. We do not restate upstream training hyperparameters here, as they are the Wan team’s and are documented in their release.

2.4 A14B mixture-of-experts: high-noise and low-noise experts

The defining feature of the Wan2.2 “A14B” models is a two-expert mixture-of-experts (MoE) design tailored to the denoising trajectory of a diffusion model. Rather than routing tokens (as in a typical LLM MoE), Wan2.2 routes *by denoising stage*, splitting the reverse process into two regimes:

1. **High-noise expert** — active during the early, high-noise steps of denoising. With little signal present, this expert focuses on the overall layout and coarse structure of the video.
2. **Low-noise expert** — active during the later, low-noise steps. With the coarse structure established, this expert refines fine details.

The handoff between experts is governed by the signal-to-noise ratio (SNR), which decreases monotonically as the denoising step index increases. The Wan team selects the switching point at a threshold tied to the SNR (described in their materials as a threshold around half of the minimum SNR), so that the high-noise expert handles the noisiest portion of the trajectory and the low-noise expert handles the rest.

Because each expert has roughly 14B parameters and only one expert is active at any given denoising step, the model totals approximately 27B parameters but keeps the per-step active parameter count — and therefore the per-step compute and memory footprint — near that of a single ≈ 14 B model. This is the same motivation that drives MoE in large language models

(increase total capacity without proportionally increasing per-step cost), applied here along the diffusion timestep axis.

2.5 Image conditioning

As an I2V model, Wan2.2-I2V-A14B conditions generation on an input image so that the generated clip begins from, and remains consistent with, that image. The output resolution and aspect ratio follow the input image dimensions within the supported 480P/720P range. An optional text prompt can additionally steer motion and content.

3 Capabilities and Reported Results

3.1 Outputs

Wan2.2-I2V-A14B generates short video clips — by default 81 frames at 16 fps (approximately five seconds) — at 480P (832×480) or 720P (1280×720), with the aspect ratio adapted to the input image. The Wan team describes the A14B MoE family as offering improved complex-motion generation and cinematic-quality output relative to the prior Wan2.1 generation.

3.2 Training-data scale (upstream-reported)

The Wan team reports that Wan2.2 was trained on substantially more data than Wan2.1: approximately **+65.6% more images** and **+83.2% more videos**. These figures are reported by the Wan team; Zen has not independently verified them.

3.3 Benchmarks (upstream-reported, attributed)

The Wan team reports that Wan2.2 achieves top performance among open- and closed-source models on their own **Wan-Bench 2.0** evaluation suite [1]. We reproduce this only as an attributed upstream claim. **Zen-Video-I2V does not contribute independent benchmark numbers**, and we deliberately omit any FVD, warping-error, temporal-consistency, or mean-opinion-score tables: prior drafts of this document contained fabricated values for such metrics, and we will not substitute new unverified numbers. Readers seeking quantitative comparisons should consult the Wan team’s published evaluations and independent third-party leaderboards.

4 Zen Integration

This section documents what Zen-Video-I2V adds on top of the upstream model. None of it constitutes a new model architecture.

4.1 Packaging

Zen-Video-I2V wraps Wan2.2-I2V-A14B with:

- Consistent command-line and API entry points aligned with the rest of the Zen stack.
- Pinned, reproducible inference defaults (resolution, frame count, fps, sampling steps, guidance) chosen to match the upstream reference configuration.
- Optional reduced-precision inference (e.g. FP8 weights) for lower-memory deployment, following community-standard quantization of the Wan weights; this is an inference-time packaging choice and does not alter the trained model.
- Provenance and safety wiring (Section 6).

4.2 Inference characteristics

Inference cost is governed by the upstream A14B architecture: per denoising step, a single $\approx 14\text{B}$ expert is active, so the compute and memory profile is close to a dense $\approx 14\text{B}$ video DiT rather than a dense 27B one. Actual latency and memory depend on resolution, frame count, sampling-step count, precision, and hardware. Because Zen-Video-I2V does not change the model, we direct users to upstream and community measurements for hardware sizing rather than publishing our own performance tables here.

4.3 API interface

Zen-Video-I2V exposes parameters consistent with the I2V task:

- **image**: input image to be animated (the fixed first frame).
- **prompt**: optional text prompt to steer motion and content.
- **resolution**: 480p (832×480) or 720p (1280×720); aspect ratio follows the input image.
- **num_frames**: number of frames to generate (default 81).
- **fps**: output frame rate (default 16).
- **steps**: number of sampling steps.
- **guidance**: classifier-free guidance scale.
- **seed**: reproducibility seed for deterministic generation.

5 Licensing

The upstream model is released under the **Apache-2.0** license by the Alibaba Wan team [2, 1]. Apache-2.0 is a permissive license that allows commercial use, modification, and redistribution, subject to its notice and attribution requirements (preserving copyright and license notices, and stating significant changes). Zen-Video-I2V is distributed consistently with these terms: we retain the upstream license and notices, attribute the work to the Alibaba Wan team, and document our integration changes.

Because the upstream license is genuinely permissive, Zen-Video-I2V can be redistributed and used commercially without the geographic or usage caps attached to some other restrictively-licensed video models. We nonetheless apply the provenance and safety measures described below as a matter of policy.

6 Provenance and Safety

6.1 Content provenance

Zen-Video-I2V emits C2PA-compatible content-provenance metadata marking output as AI-generated, in a tamper-evident manifest. This supports disclosure obligations (e.g. AI-generated-content labeling requirements) and downstream verification. Provenance emission is part of the Zen integration, not the upstream model.

6.2 Identity and consent

For inputs containing recognizable people, Zen-Video-I2V’s deployment policy requires that the caller affirm they have the rights to animate the depicted individual, applies provenance metadata to all output, and supports audit logging for accounts generating person-containing animations. These are deployment-layer controls; the underlying generative model is the upstream Wan2.2-I2V-A14B.

6.3 Misuse mitigation

Generated clips are marked as synthetic via the provenance metadata above so that downstream consumers can detect AI-generated origin. As with any open generative model, these measures reduce but do not eliminate misuse risk; responsible-use policy and access controls remain necessary.

7 Related Work

Image-to-video generation has progressed through GAN-based [5], flow-based [6], and diffusion-based [7, 8] approaches. Latent video diffusion with transformer backbones [3, 8] and large-scale video DiTs underpin the current generation of high-quality models. Wan2.2 [1, 2] contributes a denoising-stage mixture-of-experts (high-noise / low-noise experts) on top of a flow-matching latent video DiT with a 3D causal VAE. Zen-Video-I2V does not introduce new modeling techniques; it packages Wan2.2-I2V-A14B for use within the Zen stack.

8 Conclusion

Zen-Video-I2V is an Apache-2.0 redistribution and integration of Alibaba Wan’s Wan2.2-I2V-A14B image-to-video model. The model is a flow-matching latent diffusion transformer over a 3D causal VAE ($4 \times 8 \times 8$ compression), with a two-expert (high-noise / low-noise) mixture-of-experts design totaling roughly 27B parameters while keeping about 14B active per denoising step. All architectural and training credit belongs to the Wan team. Zen’s contribution is packaging: consistent tooling and inference defaults, content provenance, and safety wiring. We attribute all quantitative claims to upstream and intentionally publish no independent or fabricated benchmark numbers.

References

- [1] Wan Team, Alibaba. *Wan2.2: Open and Advanced Large-Scale Video Generative Models*. GitHub repository. <https://github.com/Wan-Video/Wan2.2>
- [2] Wan Team, Alibaba. *Wan2.2-I2V-A14B* (model card, Apache-2.0). Hugging Face. <https://huggingface.co/Wan-AI/Wan2.2-I2V-A14B>
- [3] Peebles, W. & Xie, S. (2023). Scalable Diffusion Models with Transformers. ICCV 2023.
- [4] Chung, H.W. et al. (2023). UniMax / umT5: Multilingual Pretraining for the Masses. ICLR 2023.
- [5] Vondrick, C. et al. (2016). Generating Videos with Scene Dynamics. NeurIPS 2016.
- [6] Holynski, A. et al. (2021). Animating Pictures with Eulerian Motion Fields. CVPR 2021.
- [7] Chai, L. et al. (2023). StableVideo: Text-driven Consistency-aware Diffusion Video Editing. ICCV 2023.

- [8] Blattmann, A. et al. (2023). Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127.