

Zen Live-Dub: A Permissively Licensed, Real-Time, Multi-Speaker Cross-Lingual Video Dubbing System on Commodity GPUs

Technical Report v2026.06

Zen LM Research Team
research@zenlm.org

June 2026

Abstract

We present Zen Live-Dub, a production video-dubbing system that translates speech into another language while preserving each speaker’s voice identity and lip synchronization, assembled entirely from permissively-licensed components and running in real time across two commodity machines. The work makes five contributions, each backed by direct measurement on a single NVIDIA GB10 (Grace-Blackwell, 121 GB unified memory) and an AMD “Strix Halo” (Ryzen AI Max, 16 Zen5 cores). First, a systematic *code-versus-weights* license audit that disqualifies several state-of-the-art clone-TTS models for commercial use (IndexTTS-2: non-commercial weights; F5-TTS: CC-BY-NC weights via the Emilia training set) and identifies Apache-2.0 alternatives. Second, a native lip-sync pipeline reaching **57 fps** once we localize the bottleneck to per-frame PNG I/O rather than the diffusion UNet. Third, an FP4 kernel analysis showing eager `im2col` convolution is *counterproductive* ($0.054\times$ versus `f16`) while fused implicit-GEMM offers a $\sim 4\times$ ceiling. Fourth, a multi-speaker governed pipeline—diarization, a voice registry, a consent ledger, audio watermarking, and C2PA provenance. Fifth, a distributed two-node design that breaks the single-GPU throughput wall by offloading part of the text-to-speech (TTS) workload to a second machine over direct Ethernet, achieving ≤ 8 s P95 sustained live latency with zero drops and *no new hardware*. Throughout we follow an adversarial, measurement-first methodology and report honest limitations.

1 Introduction

Automatic video dubbing—translating spoken content into a target language while keeping the speaker’s voice and lip motion—is conventionally built as a cascade: speech recognition, machine translation, voice-cloning text-to-speech, and lip synchronization. Three problems block such a cascade from becoming a licensable product, and they are rarely addressed together:

1. **Licensing.** The strongest open models frequently ship code under a permissive license but *weights* under a non-commercial one, an asymmetry that silently disqualifies them for a commercial deployment.
2. **Real-time cost.** Diffusion lip-sync and autoregressive TTS are assumed to require data-center accelerators; the actual bottlenecks are often elsewhere.
3. **Governance.** A newsroom cannot ship synthetic voices without consent tracking, provenance, and disclosure that satisfy emerging law.

Zen Live-Dub addresses all three on commodity hardware. We report what we measured, including results that contradicted our own initial assumptions, and we describe the adversarial development methodology that surfaced those contradictions.

2 System Architecture

The pipeline is a directed stage graph; each stage is a swappable component behind a stable interface:

ingest $\rightarrow$ diarize $\rightarrow$ registry+consent $\rightarrow$ S2TT $\rightarrow$ clone-TTS $\rightarrow$ isochrony $\rightarrow$ [lip-sync] $\rightarrow$ remix $\rightarrow$ watermark $\rightarrow$

Speaker identity is not stored as a separate model per voice; it is a small embedding in a vector index. A registry of N enrolled voices costs N embeddings plus reference clips behind a *single* TTS model, so the registry size is a vector-search problem rather than a GPU-memory problem. Concurrent speakers in one stream are capped at four by streaming diarization; offline diarization is unbounded.

3 Real-Time Lip Synchronization

3.1 The bottleneck is I/O, not the kernel

Our lip-sync stage uses a Stable-Diffusion-derived UNet to regenerate the mouth region conditioned on the dubbed audio. The UNet alone runs at 152fps (f16, batch 8) in isolation, yet the end-to-end pipeline initially ran at 4fps. Profiling the production pipeline revealed it was executing in fp32 and—more importantly—spending 12.9 ms/frame writing per-frame PNG files and 4.6 ms/frame reading them back. Replacing per-frame PNG encode/read with a persistent raw-video pipe and a one-time frame preload, under fp16 with a tiny VAE decoder (TAESD), isolates the gain cleanly (Table 1).

Table 1: Lip-sync I/O fix, A/B at identical model path (sun6, cached avatar, bs8)

Configuration	as-built ms/frame	e2e fps	model-path fps
Legacy (per-frame PNG)	40.99	24.40	53.05
Raw-video pipe + preload	21.71	46.06	53.69
Pipe, sustained (5125 frames)	17.54	57.02	70.51

The model path is identical across configurations (~ 53 fps), so the entire +21.7fps delta is the I/O change. The 57fps figure was verified by independently counting 5125 encoded frames and cross-checking wall-clock (84.9 s) against process timing (89 s). The headline lesson: the exotic kernel was never on the critical path to real-time lip-sync.

3.2 FP4 convolution: a disproven shortcut

We tested whether NVFP4 (4-bit) convolution via `im2col` + `_scaled_mm` accelerates the UNet on Blackwell. It does not, when activation quantization and `im2col` are honestly inside the timed region (Table 2). The matrix-multiply primitive is genuinely fast ($2.6\text{--}4.4\times$ versus bf16), but the surrounding cost dominates: dynamic activation quantization is 90% of runtime and explicit `im2col` alone exceeds the entire cuDNN f16 convolution stack.

Table 2: Eager FP4 convolution versus f16 (whole UNet conv stack, batch 8)

Path	Time (ms)	Speedup vs f16
cuDNN f16 (baseline)	25.5	1.00×
Eager NVFP4 <code>im2col+_scaled_mm</code>	470.4	0.054 ×
– of which activation quant	422.7	—
– of which <code>im2col</code>	34.4	—
– of which FP4 GEMM	6.2	—
Pure GEMM ceiling (conv-equiv shapes)	—	2.6–4.4×

Numerics are sound (per-shape cosine ≥ 0.9907 across all 35 UNet shapes) *only* when block scales use the cuBLAS `to_blocked` 128×4 swizzle (output cosine 0.999997); the naive padded layout produces numerical garbage (cosine 0.18). The actionable conclusion is that a real FP4 win requires a *fused* implicit-GEMM convolution—quantization in the mainloop, scales emitted in the swizzled layout—not an eager decomposition.

4 Permissively Licensed Component Selection

Selecting a commercially-usable clone-TTS is gated by the weights license, not capability. We audited the 2025–2026 field against primary sources (model-card metadata, training-set licenses, technical reports), distinguishing the license of the *code* from that of the *weights*, and the ability to clone an *arbitrary* speaker from fixed voice packs (Table 3).

Table 3: Clone-TTS license and capability audit (weights, not code)

Model	Weights license	Arbitrary clone	Verdict
Qwen3-TTS (open)	Apache-2.0	yes	adopt
CosyVoice 3 (0.5B)	Apache-2.0	yes	backup
IndexTTS-2	Bilibili (non-comm.)	yes	reject
F5-TTS	CC-BY-NC-4.0 (Emilia)	yes	reject
Kokoro-82M	Apache-2.0	no (fixed voices)	reject
Qwen3-TTS-Flash	none (hosted API)	yes	reject

Two traps recur. The *code-versus-weights* trap: F5-TTS code is MIT but its checkpoint inherits CC-BY-NC from the Emilia dataset, and the maintainer confirms that fine-tuning does not relicense it. The *fixed-voice* trap: Kokoro is Apache-2.0 but offers a closed set of voice packs and cannot clone a target speaker. The same audit applies to lip-sync: the diffusion UNet weights carry a permissive (OpenRAIL-M) license, but the conventional blend mask depends on a BiSeNet face-parser trained on the non-commercial CelebAMask-HQ. We replace it with a FAN-2D-4 landmark blend (BSD), removing the only hard non-commercial dependency in the visual path while preserving quality (Section 6).

5 Voice Cloning

The adopted backbone, Qwen3-TTS (open, Apache-2.0), performs zero-shot cross-lingual cloning from a short reference (Table 4). The shipping use case—a Chinese reference producing English output in the same voice—is the demanding “cross-lingual” row; identity is measured as speaker-embedding cosine and intelligibility as round-trip word error rate (WER).

Table 4: Cross-lingual zero-shot clone (zh reference $\rightarrow$ target), speaker cosine

Engine	cross-ling. cosine	WER	license
Qwen3-TTS (open)	0.871	~ 0.0	Apache-2.0
CosyVoice 3	0.835	0.0	Apache-2.0
IndexTTS-2 (disqualified)	0.935	0.02	non-commercial

The license-disqualified IndexTTS-2 leads on raw similarity, but it cannot ship; Qwen3-TTS is the best *usable* option and additionally repairs a coverage failure: IndexTTS-2’s tokenizer cannot encode accented Latin characters, emitting nothing for Spanish, French, and German, whereas Qwen3-TTS supports ten languages natively (English 0.924, French 0.935, German 0.902; Spanish 0.888, intelligible but below a 0.90 gate, motivating a 0.88 newsroom floor).

Multi-speaker handling is a pipeline, not a model: streaming diarization tags who-speaks-when, each segment is matched against the registry (identification accuracy $\geq 95\%$ on the enrolled set; 0.01 ms per query, scaling to thousands of entries), and per-speaker references drive the clone. Source separation (Demucs) isolates speech from the music/SFX bed so the dub is remixed under the preserved background (+6 dB SI-SDR; 74% of bed energy retained).

6 Governance, Provenance, and the Permissively Licensed Visual Path

A newsroom deployment must satisfy consent and disclosure law (e.g. the Tennessee ELVIS Act, in force; the EU AI Act Article 50 synthetic-audio disclosure requirement, effective August 2026). The pipeline enforces a signed, revocable consent record per governed voice, checked at synthesis time; an unconsented speaker is refused and routed to a silent hold. Provenance is a C2PA manifest whose `consent_ref` foreign-key is verified coherent with the consent ledger. Synthetic-audio watermarking uses AudioSeal (MIT), which in our tests detected at 100% (zero bit-error, zero false-positive) across clean, MP3-128k, AAC-128k, Opus-64k, and double-encoded chains, at 33.1 dB SNR (PESQ 4.52). The end-to-end governed run passed 20/20 assertions on GPU.

For the visual path, the FAN-landmark blend that removes the BiSeNet/CelebAMask-HQ dependency is statistically indistinguishable from the original: lip-sync error LSE-C = 6.924 at audio-video offset -2 , within 0.005 of the BiSeNet baseline (6.93), with composite fidelity 47.85 dB PSNR / 0.993 SSIM. The visual path is therefore license-clean with no measurable quality cost.

7 Distributed Two-Node Real-Time Dubbing

7.1 The single-GPU wall is utilization, not speed

For live operation we target the latency budget of broadcast, not of conversation: live television runs a 5–10 s safety delay and captions average ~ 5.6 s, so a 3–8 s speech-in-to-dubbed-speech-out window is acceptable. On one GB10, the streaming orchestrator is correct (bounded queues, in-order release, graceful drop-oldest under overload) but cannot meet the budget: a single TTS worker pinned at $\sim 90\%$ busy accumulates head-of-line queue residency, pushing end-to-end P50 to 37.6 s. Critically, a *second* TTS worker on the *same* GPU does not help—the two contend for the same compute, doubling per-unit service time for no net throughput.

7.2 Offloading TTS to a second machine

The constraint is throughput on one accelerator, and the data exchanged (16 kHz mono audio plus a reference clip) is tiny. We therefore place a second TTS node on a separate machine connected by direct Ethernet (measured: 0.41 ms RTT, 2.15 Gbit/s; per-unit transport overhead 4–6 ms). The second node is an AMD Strix Halo whose GPU path is blocked by a non-CUDA tooling gap (the DirectML backend rejects the model’s fused `GroupQueryAttention` op), so it runs a permissively-licensed clone-TTS (Chatterbox-ONNX) on its 16 CPU cores at real-time factor 0.94–0.97. Because the two nodes do not share an accelerator, the contention vanishes (Table 5).

Table 5: Live streaming latency, 3.2-minute sustained run

Configuration	P50 (s)	P95 (s)	drops	spark util	evo util
1 node (GB10)	37.6	—	—	~90%	—
2 workers, 1 GPU	54.8	—	6	191%	—
2 nodes (GB10 + Strix Halo)	4.04	6.14	0	68%	63%

A balanced split (~53% local GPU / 47% remote CPU) with short commit units drops each node below 90% busy; queue residency collapses and the entire latency distribution falls under the 8 s budget (max 7.92 s) with zero drops over 3.2 minutes, while per-speaker identity is preserved (cosine 0.866/0.878 across nodes). Aggregate throughput is 2.19× real time. We explicitly tested and rejected a three-slot configuration: a second CPU worker oversubscribes the 16-core machine, inflating tail latency. The result is real-time dubbing on two existing commodity machines with no new hardware; the throughput law is

$$\text{sustained rate} = \frac{1}{\max_i \text{service}_i}, \quad \text{e2e latency} = \sum_i \text{service}_i + \text{queue} + \text{jitter},$$

and distributing the dominant stage across non-shared resources is what makes both quantities fit.

8 Methodology: Adversarial, Measurement-First Development

The system was built and validated by a fleet of cooperating agents under a discipline of *honesty over success*: build, then independently red-team, then prove on hardware, with claims rejected unless backed by a measured number. This methodology repeatedly caught errors that a success-oriented process would have shipped: a re-blend harness that reported “LSE-C parity” while in fact desyncing audio by 15 frames (caught by calibrating the evaluator against a known-good clip); a headline “57 fps over 5125 frames” that was real but whose frame count required independent recounting because an intermediate file had been trimmed; and a composite-fidelity number that had been measured but never persisted to disk. We regard the disproven results (eager FP4 convolution; single-GPU live real-time) as first-class contributions: a cleanly measured negative result is what redirected effort to the changes that worked.

9 Limitations

We report the following honestly. (1) Spanish clone similarity (0.888) sits just under a 0.90 gate; it is intelligible but motivates either a 0.88 editorial-review floor or longer enrolled references. (2) Per-1 s-fragment speaker cosine reads ~0.74 on short streaming units (a d-vector stability artifact, equal on both nodes); the continuous-track perceptual measure clears 0.85, but tightening the fragment measure trades against the latency budget. (3) More than four concurrent live

speakers (overlapping crosstalk) remains unsolved industry-wide; we degrade gracefully rather than fail silently. (4) The second node runs on CPU because the AMD GPU path is blocked by the DirectML `GroupQueryAttention` gap; a faster path (a GQA-decomposed ONNX export, or ROCm on native Linux) would add headroom beyond the 8 s budget. (5) Quality is reported on a limited clip family; broad multi-face, multi-language, and long-duration generalization is future work.

10 Conclusion

Zen Live-Dub demonstrates that a license-clean, multi-speaker, cross-lingual video dub—offline and live—is achievable on commodity hardware. The decisive engineering insights were not the expected ones: real-time lip-sync was gated by per-frame I/O rather than the diffusion kernel; commercial viability was gated by weights licenses rather than model quality; and live real-time was gated by single-accelerator utilization rather than raw speed, broken by offloading the dominant stage to a second machine over a direct link. Every shipping component is permissively licensed and every headline number was independently verified. The remaining gaps are throughput headroom and quality generalization, not architecture or license.

References

- [1] Zhang, Y. et al. MuseTalk: Real-Time High-Quality Lip Synchronization with Latent Space Inpainting. *arXiv:2410.10122*, 2024.
- [2] Li, C. et al. LatentSync: Audio Conditioned Latent Diffusion Models for Lip Sync. *arXiv:2412.09262*, 2024.
- [3] Du, Z. et al. CosyVoice 3: Towards In-the-Wild Speech Generation. *arXiv:2505.17589*, 2025.
- [4] Qwen Team. Qwen3-TTS Technical Report. *arXiv:2601.15621*, 2026.
- [5] San Roman, R. et al. Proactive Detection of Voice Cloning with Localized Watermarking (AudioSeal). *ICML*, 2024.
- [6] Coalition for Content Provenance and Authenticity. C2PA Technical Specification v2.2, 2025.
- [7] Park, T. et al. Sortformer: Streaming Speaker Diarization. *arXiv:2409.06656*, 2024.
- [8] NVIDIA. NVFP4 and the Blackwell Tensor Core Microscaling Formats. Technical Documentation, 2025.
- [9] He, H. et al. Emilia: A Large-Scale, In-the-Wild Multilingual Speech Dataset. *arXiv:2407.05361*, 2024.
- [10] European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 50, 2024.