60-second takeaway
We ran a consistent single-speaker benchmark on four open-source TTS models using IMDA NSC FEMALE_01 on an RTX 3090 Ti (24GB). VoxCPM 1.5 and Qwen3-TTS 1.7B both produced deployable outputs. IndexTTS2 gave a stable full-SFT baseline. CosyVoice3 finetuning did not reach production quality in this run (rerun pending).
If you need something deployable today on a 24GB GPU, start with VoxCPM or Qwen3-TTS LoRA.

What this benchmark covers

This is a practitioner-oriented comparison, not an academic leaderboard. We evaluated four models under the same conditions:

- **Dataset:** IMDA NSC FEMALE_01 - a single-speaker set with natural Singaporean English accent
- **Hardware:** one NVIDIA RTX 3090 Ti (24 GB VRAM)
- **Goal:** produce voice-cloned audio suitable for AI-generated video narration (A-roll use case)
- **Evaluation:** qualitative listening on naturalness, long-text stability, accent retention, and operational friction

We are not measuring WER or MOS scores from automated tools. We are measuring whether the output sounds production-ready to a human listener on a video platform.

The four models

VoxCPM 1.5

VoxCPM 1.5 uses a LoRA finetuning path that fits within 24GB VRAM without modification. Training is straightforward with standard train/val splits.

Dimension	Result
Finetuning approach	LoRA
Best checkpoint (this run)	step_0004000
Long-text stability	Good

Prompt sensitivity	Moderate - use clean prompt clips
Production-ready?	**Yes**

Key insight: No-prompt generation at step 4000 gave the best naturalness. Prompted inference copied prompt room noise into the output, which was audible on studio playback. Use prompt only when strong speaker lock is required.

Qwen3-TTS 1.7B

Qwen3-TTS 1.7B with LoRA was the model where adapter scale mattered most. Scale 1.0 over-steered and produced noisy outputs; scale 0.3 to 0.35 sounded stable.

Dimension	Result
Finetuning approach	LoRA
Best checkpoint (this run)	Epoch 10
Best LoRA scale	0.3 to 0.35
Long-text stability	Good with SDPA backend
Prompt sensitivity	Low - robust to formatting variation
Production-ready?	**Yes**

Key insight: The scale sweep matters more than checkpoint selection alone. Run a quick 5-sample listening test at scales 0.2, 0.3, 0.35, and 0.5 before committing to a checkpoint. Scale 1.0 is almost always wrong for this benchmark.

IndexTTS2

IndexTTS2 uses full SFT (not LoRA). It requires more careful checkpoint management because the training loop had crash recovery issues in our run.

Dimension	Result
Finetuning approach	Full SFT
Best checkpoint (this run)	model_step14000.pth
Long-text stability	Good
Crash recovery	Required explicit resume management
Production-ready?	**Yes - with operational caution**

Key insight: Keep ALL checkpoints until you've done a listening eval sweep. The retention policy deleted older checkpoints before we could test them. Pin the best checkpoint explicitly once identified - don't rely on automatic deletion logic.

CosyVoice3

CosyVoice3 LoRA finetuning was the outlier. Our first run did not reach production quality.

Dimension	Result
Finetuning approach	Full SFT via LoRA path
Run status	Not production-ready (this run)
Main failure modes	Checkpoint drift, long-text instability, prompt sensitivity
Production-ready?	**No - rerun pending**

CosyVoice2 baseline (zero-shot, no finetuning) sounded acceptable as a control reference. See CosyVoice LoRA Fine-Tuning: What Worked, What Didn't, and the Rerun Plan for the full diagnostics and rerun plan.

Head-to-head comparison

Model	Finetuning	VRAM (24GB)	Best checkpoint	Deployable now?	Effort to run
VoxCPM 1.5	LoRA	Fits	step 4000	Yes	Low
Qwen3-TTS 1.7B	LoRA	Fits	Epoch 10, scale 0.3	Yes	Low–Medium
IndexTTS2	Full SFT	Fits	step 14000	Yes (with care)	Medium
CosyVoice3	Full SFT (LoRA path)	Fits	Rerun pending	Not yet	High

Decision guide

If you need deployable output fastest

Start with **VoxCPM 1.5 step 4000**. It had the lowest setup friction and the cleanest no-prompt output in our run. LoRA training is straightforward and the checkpoint selection rule is simple.

If you need LoRA-style adapter control

Use **Qwen3-TTS 1.7B LoRA**. The scale parameter gives you a post-training knob to tune output strength without retraining. This is valuable when you want to fine-tune the output on different content types without full retraining cycles.

If you need the most reproducible full-SFT baseline

Use **IndexTTS2**. Full SFT converges more predictably than LoRA for some voice profiles. The crash recovery requirement is manageable once you have an explicit checkpoint retention policy.

If you want to evaluate CosyVoice

Use **CosyVoice2 as a zero-shot baseline** while waiting for the CosyVoice3 rerun. Do not deploy the current CosyVoice3 run.

What is IMDA NSC FEMALE_01?

IMDA NSC is the National Speech Corpus published by Singapore's Infocomm Media Development Authority. FEMALE_01 is a single-speaker subset with natural Singaporean English. We use it as a benchmark voice because it has a distinctive accent profile that stress-tests speaker similarity in voice cloning - a model that sounds natural on this speaker generalises well to other non-American-English speakers.

Audio evidence

All audio samples from this benchmark are published in the individual model deep dives. Listen to them side by side before making a deployment decision.

- VoxCPM audio samples
- Qwen3-TTS audio samples
- IndexTTS2 audio samples
- CosyVoice audio samples

FAQ

Can I run all four models on a single RTX 3090 Ti (24GB)?

Yes. All four models fit within 24GB VRAM for both training and inference. The full feasibility notes - including peak VRAM, runtime, and recipe availability - are covered in Voice Cloning on a 24GB GPU: What Actually Works in 2026.

Which model has the best Singaporean English accent retention?

In this benchmark, VoxCPM and IndexTTS2 both retained the FEMALE_01 accent profile well. Qwen3-TTS at the right scale also retained it. CosyVoice3 (current run) had inconsistent retention.

Are any of these models commercially licensed for production use?

License status varies. IndexTTS2 uses a research license with commercial use restrictions. VoxCPM, Qwen3-TTS, and CosyVoice have varying commercial terms - verify the latest license on each model's repository before deploying.

What's the difference between LoRA and full SFT for TTS finetuning?

LoRA (Low-Rank Adaptation) trains a small adapter on top of a frozen base model. It uses less VRAM and trains faster, but the adapter's strength needs tuning (the `lora_scale` parameter). Full SFT (Supervised Fine-Tuning) updates all model weights. It requires more VRAM and longer training but tends to converge more reliably for voice profiles with strong accent characteristics.

Sources

- Full benchmark matrix: IMDA NSC Voice Cloning Finetuning Benchmark 2026
- Evidence map and artifact paths: `reports/tts-experiments-evidence-map.md`
- VoxCPM deep dive: VoxCPM 1.5 LoRA Finetuning on IMDA NSC FEMALE_01
- Qwen3-TTS deep dive: Qwen3-TTS LoRA Fine-Tuning: Scale Sweeps, Checkpoints, and Production Defaults
- IndexTTS2 deep dive: IndexTTS2 Finetuning on IMDA NSC FEMALE_01
- CosyVoice diagnostics: CosyVoice LoRA Fine-Tuning: What Worked, What Didn't, and the Rerun Plan