60-second takeaway
We ran one consistent single-speaker benchmark on IMDA NSC FEMALE_01 with a single-GPU setup.
VoxCPM, IndexTTS2, and Qwen3-TTS all produced usable outputs under specific settings; CosyVoice3 did not reach production-ready quality in this run.
Treat this as an execution benchmark under one configuration, not a universal model ranking.

Who this is for

- **Founder / strategy reader:** use the matrix and decision guide to pick what to deploy next.
- **Engineer reader:** use each linked deep dive for exact recipes, checkpoints, and failure diagnostics.

Shared experiment setup

- **Dataset:** IMDA NSC single-speaker set (`FEMALE_01`), with model-specific preprocessing.
- **Hardware:** single NVIDIA RTX 3090 Ti (24 GB VRAM).
- **Evaluation:** qualitative listening on naturalness, accent retention, noise profile, long-text stability, and operational friction (VRAM, disk, rerun complexity).

Comparison matrix

Model	Dataset handling	Train recipe	Best checkpoint in this run	Main failure mode	Recommended inference setting
CosyVoice2 (baseline/control)	Baseline sample used as control	No finetune in this benchmark	Baseline control sample only	Not evaluated as a finetune target in this series	Use as control reference only
CosyVoice3	IMDA NSC `FEMALE_01` via Cosy pipeline	Full SFT (0.5B path)	Early epochs were better than later ones	Unstable long-form behavior and weak	Strict prompt formatting and conservative

				linguistic consistency	checkpoint selection
VoxCPM 1.5	`FEMALE_01` resampled to 44.1 kHz + train/val split	LoRA finetune to step 9000	`step_0004000` (lowest val total in run)	Prompted outputs can inherit prompt noise strongly	For naturalness, no-prompt run at step 4000; for speaker lock, use clean prompt clip
Qwen3-TTS 1.7B + LoRA	`FEMALE_01_44k` JSONL + codec prep	LoRA train/val/test split runs	Epoch 10	LoRA can over-steer at scale 1.0 and create noisy outputs	`lora_scale` around 0.3, SDPA backend, deterministic decode
IndexTTS2	`FEMALE_01_44k` processed manifests	Full SFT with resumes	`model_step14000.pth`	Training interruptions and crash-prone resume path	Keep stable resume strategy, compare 14000 vs 15949 by listening

CosyVoice3 status: Not production-ready (current run)

Interpretation note: "Not production-ready" here means "not production-ready in this experiment configuration." We plan a follow-up CosyVoice3 rerun with revised settings.

Decision guide

If you need deployable output fastest

Start with **VoxCPM step 4000** or **Qwen3-TTS LoRA epoch 10 at scale 0.3 to 0.35**.

If your priority is configuration stability and reproducibility

Use **IndexTTS2** as a stable full-SFT reference and keep checkpoints pinned by explicit listening tests.

If you are evaluating CosyVoice for this dataset

Use **CosyVoice2 as your baseline control** and treat **CosyVoice3 in this run** as a diagnostics case, not a final verdict.

Deep dives

- **VoxCPM:** https://instavar.com/blog/VoxCPM_1_5_LoRA_Finetuning_IMDA_NSC_FEMALE_01
- **IndexTTS2:** https://instavar.com/blog/IndexTTS2_Finetuning_IMDA_NSC_FEMALE_01
- **Qwen3-TTS:** https://instavar.com/blog/LoRA_Finetuning_Qwen3_TTS_Custom_Voices
- **CosyVoice2 vs CosyVoice3:** https://instavar.com/blog/CosyVoice2_vs_CosyVoice3_IMDA_NSC_FEMALE_01

Evidence and reproducibility

Full artifact mapping (checkpoints, sample paths, logs):

- `reports/tts-experiments-evidence-map.md`

If you want the short version for exec stakeholders, use this page. If you want exact commands and failure traces, open the model-specific deep dives above.