60-second takeaway
VoxCPM produced a strong practical result in this benchmark once we stabilized dataset prep and used a clean inference protocol.
The best balance of quality and stability came from `step_0004000` in our run.
Prompted inference can over-copy prompt noise, so prompt quality matters as much as checkpoint choice.

Where this fits

- **For founders:** VoxCPM is a viable production candidate from this benchmark.
- **For engineers:** use this page for train recipe, checkpoint pick logic, and inference defaults.

For the series overview matrix, see:

- https://instavar.com/blog/IMDA_NSC_Voice_Cloning_Finetuning_Benchmark_2026

Experiment setup

- **Base model:** VoxCPM1.5
- **Dataset:** IMDA NSC `FEMALE_01`
- **Audio prep:** resampled to 44.1 kHz for VoxCPM1.5 path
- **Hardware:** RTX 3090 Ti 24 GB
- **Training mode:** LoRA fine-tuning

Best checkpoint logic

We tracked validation total loss across steps and selected the strongest zone by both trend and listening:

- Best recorded validation total in this run was at `step_0004000`.
- Later checkpoints remained usable, but were not consistently better on subjective naturalness.

Audio evidence

Best practical sample (this run)

`Settings`: no-prompt, no denoiser, long text test.

Failure modes we saw

- Prompted outputs can inherit prompt-room noise strongly.
- Denoisers can clean hiss but also shift timbre and bandwidth perception.
- Long-form outputs are sensitive to prompt clip quality and consistency.

Recommended inference settings

For this exact benchmark setup:

- Start from `step_0004000` as default checkpoint.
- Use no-prompt generation first to estimate model prior naturalness.
- Add prompt only when you need stronger speaker lock.
- Use denoiser only when hiss/noise is clearly audible.

Engineer appendix

Key paths from this run

- Checkpoints: `/mnt/work/chee-wei-jie/voice-model-outputs/voxcpm/female01/ckpts`
- Analysis samples: `/mnt/work/chee-wei-jie/voice-models/VoxCPM/analysis/ab_samples`

Notes

- Validation and sample interpretation were paired; we did not choose by scalar loss only.
- Metadata and evidence mapping are recorded in `reports/tts-experiments-evidence-map.md`.

Related deep dives

- [Qwen3-TTS LoRA]()

- IndexTTS2
- CosyVoice2 vs CosyVoice3