**60-second takeaway**
CosyVoice3 LoRA fine-tuning on IMDA NSC FEMALE_01 did not reach production quality in our first run. CosyVoice2 baseline audio was acceptable as a control.
The failure was configuration-specific: checkpoint drift, prompt sensitivity, and operational fragility in the training loop.
We have a clear rerun plan with tighter checkpoint gating and a fixed prompt harness - this post documents what to reproduce and what to change.

## Where this fits

This is a run-specific engineering note in the IMDA NSC Voice Cloning Finetuning Benchmark 2026 series. The other models in that series - VoxCPM 1.5 and Qwen3-TTS LoRA - both produced deployable results. CosyVoice3 is the outlier that needs a rerun before it can be evaluated fairly.

- **For founders:** do not deploy the current CosyVoice3 run. Use VoxCPM or Qwen3-TTS while the rerun is pending.
- **For engineers:** use this page as the diagnostic handoff for the next CosyVoice LoRA run.

## CosyVoice2 vs CosyVoice3: what the benchmark found

| | CosyVoice2 (baseline/control) | CosyVoice3 (current run) |
|---|---|---|
| **Training mode** | No finetune - zero-shot reference | Full SFT via LoRA path (`train_cosyvoice3_lora.py`) |
| **Qualitative result** | Acceptable naturalness, usable as control | Did not reach production quality in this run |
| **Long-form stability** | Stable | Unstable beyond ~20 seconds |
| **Linguistic consistency** | Consistent | Weak in listening checks |

| | | |
|---|---|---|
| **Production-ready?** | Yes (zero-shot baseline) | No (current run) |

The key conclusion: this is not a claim that CosyVoice3 is inherently worse than CosyVoice2. It is a claim that our first CosyVoice3 run configuration produced a worse result, and we know why.

## Audio evidence

### CosyVoice2 baseline/control

### CosyVoice3 representative sample (this run - not production-ready)

Listen to the two clips side by side. The CosyVoice2 clip is cleaner on naturalistic prosody. The CosyVoice3 clip shows the instability in long-form decoding that blocked production deployment.

## Failure mode analysis

Three contributing factors explain the current run outcome:

### 1. Checkpoint quality drift

CosyVoice3 LoRA training showed a pattern where early epochs were noticeably better than later ones. We did not have a strict per-epoch validation gate in place, so the run continued past the best region. The practical lesson: CosyVoice3 needs tighter checkpoint gating with explicit listening checkpoints every 2–3 epochs, not just loss curve tracking.

### 2. Prompt sensitivity and long-text decoding

CosyVoice3 in this run was sensitive to prompt formatting. Small changes to how the text was segmented before inference changed output quality significantly. Long-text generation (>20s) showed linguistic inconsistency - words dropped or merged in a way that sounded unnatural on listening review. VoxCPM and Qwen3-TTS were more robust to prompt formatting variation.

### 3. Operational fragility: large checkpoint churn

The training loop produced large checkpoint files at short intervals. Without explicit retention policy, older (potentially better) checkpoints were overwritten. By the time we evaluated, some of the best early-epoch checkpoints were no longer available. This is the same issue that affected the IndexTTS2 run, but with a more acute impact here because the best zone was earlier.

# Rerun plan

The next CosyVoice LoRA run should fix all three failure modes:

```
1. Checkpoint gating
     Set explicit save_every_n_epochs = 2 (not just save_by_loss)
     Keep ALL checkpoints until listening eval confirms best region
     Do not rely on val loss alone — add listening gate before deleting

2. Prompt formatting harness
     Fix a single prompt formatting template for all comparisons
     Use the same segmentation rules as the VoxCPM/Qwen3 runs
     Test 5s, 15s, and 30s generation lengths at each checkpoint gate

3. Long-text stability eval
     Include a 30s+ test clip in every checkpoint gate evaluation
     Fail a checkpoint if it drops or merges words in the long clip
     Only promote a checkpoint when it passes all three length tiers
```

## Training tool reference

- Train script: /mnt/work/chee-wei-jie/voice-models/CosyVoice/tools/train_cosyvoice3_lora.py
- Inference script: /mnt/work/chee-wei-jie/voice-models/CosyVoice/tools/infer_cosyvoice3_lora.py
- Run artifacts: /mnt/work/chee-wei-jie/voice-models/CosyVoice_runs/female01_cv3_lr1e5_run1/

# When to use CosyVoice vs alternatives

While the CosyVoice3 rerun is pending, here is the practical routing:

| Situation | Recommended model |
|---|---|
| Need deployable output now on 24GB GPU | VoxCPM 1.5 (step 4000) |

| | |
|---|---|
| Need LoRA-style fine-tuning with scale control | Qwen3-TTS 1.7B (epoch 10, scale 0.3–0.35) |
| Need full SFT baseline with crash recovery | IndexTTS2 (step 14000) |
| Evaluating CosyVoice specifically | Use CosyVoice2 as control; await CosyVoice3 rerun |

For the full series comparison matrix, see IMDA NSC Voice Cloning Finetuning Benchmark 2026.

# FAQ

### Is CosyVoice LoRA fine-tuning worth trying on a 24GB GPU?

Yes - the tools exist (`train_cosyvoice3_lora.py`) and the VRAM footprint is manageable. The issue in our run was not VRAM but checkpoint management and prompt formatting. With the rerun plan above, the 24GB path should work.

### Why did CosyVoice2 baseline hold up better than the CosyVoice3 finetune?

CosyVoice2 was used as a zero-shot reference, not a finetuned model. Zero-shot inference on a well-pretrained model avoids all the failure modes in the finetuning loop (checkpoint drift, prompt sensitivity, checkpoint churn). The comparison is not fair - it's a pretrained zero-shot vs a partially-successful finetune. A well-executed CosyVoice3 finetune should outperform the CosyVoice2 zero-shot baseline.

### What learning rate was used in the run?

`lr1e5` (1e-5), from the run artifact path `female01_cv3_lr1e5_run1`. This is a standard starting point for TTS LoRA runs and is not the primary failure cause.

# Related deep dives

- VoxCPM 1.5 LoRA Finetuning on IMDA NSC FEMALE_01
- Qwen3-TTS LoRA Fine-Tuning: Scale Sweeps, Checkpoints, and Production Defaults
- IndexTTS2 Finetuning on IMDA NSC FEMALE_01
- CosyVoice 3: Zero-Shot Multilingual TTS at 1.5B Parameters