**60-second takeaway**
Qwen3-TTS + LoRA worked well on this benchmark once we controlled inference scale.
The key lesson was not just checkpoint selection but adapter strength: scale 1.0 over-steered, while 0.3 to 0.35 sounded stable.
For this run, epoch 10 plus `lora_scale` around 0.3 was the best operating point.

## Companion repo

All reusable LoRA tooling is published separately:

- https://github.com/cheeweijie/qwen3-tts-lora-finetuning
- https://github.com/cheeweijie/qwen3-tts-lora-finetuning/releases/tag/v0.1.1

## Where this fits

- **For founders:** this is a strong candidate if you want high quality from single-GPU LoRA runs.
- **For engineers:** this page captures exact run behavior, including where losses flattened and where inference destabilized.

Series overview:

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

## Experiment setup

- **Model:** Qwen3-TTS 1.7B Base + LoRA
- **Dataset:** IMDA NSC `FEMALE_01_44k`, JSONL + codec prep pipeline
- **Split:** train/val/test = 90/5/5
- **Hardware:** RTX 3090 Ti 24 GB

## Best checkpoint logic

- Validation improved early and flattened around epochs 8 to 12.

- Validation started rising after epoch 13 in our continued run.
- Best checkpoint by validation trend in this run: **epoch 10**.

# Audio evidence

## Recommended sample from this run

`Settings:` epoch 10 adapter, scale 0.35.

# Failure modes we saw

- Scale 1.0 often sounded noisy/over-steered.
- Some background inference runs failed due to environment/runtime issues, not model quality.
- Attention backend choice affected inference stability in long sweeps.

# Recommended inference settings

For this run and hardware profile:

- Use epoch 10 adapter as the first candidate.
- Set `lora_scale` to `0.3` by default (safe range: `0.25` to `0.35`).
- Use `attn_implementation=sdpa` if Flash-Attention path is unstable.
- Prefer deterministic decode for clean A/B comparisons.

# Engineer appendix

## Key paths from this run

- Local LoRA companion repo: `/mnt/work/chee-wei-jie/voice-models/qwen3-tts-lora-finetuning`
- Qwen source workspace: `/mnt/work/chee-wei-jie/voice-models/Qwen3-TTS`

## Distribution note

The companion repo approach keeps upstream Qwen3-TTS clean while allowing pinned, reproducible patches and scripts.

## Related deep dives

- VoxCPM
- IndexTTS2
- CosyVoice2 vs CosyVoice3