Group — Audio Source Separation (Vocals ↔ Accompaniment)

*Run **3 – 4 tightly-scoped experiments** in one free-tier-Colab week (≈ 8 GPU-h) that show **what actually improves separation quality** on MUSDB18-HQ. Getting perfect karaoke tracks is not required; understanding the trade-offs is.*

---

0 First-evening fixes & sanity checks

Why it matters	Quick remedy
Time-domain baseline hides phase issues	Move to **STFT masking** ASAP: feed magnitude, predict *soft mask* ∈ [0, 1], reuse mixture phase for iSTFT. You already discovered phase is hard .
Tiny batch, high variance	In Colab T4 you can fit **batch = 8 clips × 6 s × 44.1 kHz** with 1.3 GB VRAM if you keep tensors in `float16`.
**Segment imbalance (65 % vocal-free) **	Use **class-balanced sampler**: draw 50 % clips that contain vocals, 50 % without, or weight the loss by target RMS.
Evaluation placeholders	Install `musdb==0.4.0`, `museval==0.4.1`, `torchmetrics[audio]`. Verify that the pre-trained *Open-Unmix (umxhq)* gets ≈ 5.2 dB SDR on the MUSDB test vocals → proves the metric pipeline is correct.

---

1 Solid baseline to lock in (≤ 1 GPU-h)

Item	Setting
Model	*Open-Unmix-small* (3 bidirectional GRU layers, 384 hidden, 3 M params).
Input	STFT 1024 hop 256, segment 6 s (4096 frames).
Loss	L1 on magnitude × weighting factor $\alpha(f) = \sqrt{f}$ (gives low-freq more weight).
Optimizer	AdamW, lr 3×10^{-4}, cosine decay 20 epochs, early-stop patience = 3.
Hardware	Mixed precision, batch 8 → 35 min per 20-epoch run on T4.

Log SDR, SI-SDR, SAR on the 50 validation tracks; these become the numbers every later run must beat.

---

2 Experiment menu — pick any three

ID	Hypothesis	Change vs. baseline	Expected gain	GPU h
A	**Demucs-tiny** (time-domain conv-transposedconv) captures transients better than spectrogram masking.	`facebookresearch/demucs` “dnb” config, 3 layers, 64 ch; train 15 epochs.	↑ SI-SDR on drums/bass (+1 dB)	2

ID	Hypothesis	Change vs. baseline	Expected gain	GPU h
B	**Multi-task mask prediction** (vocals *and* accompaniment) stabilises training.	Single net, 2-head output; loss = L1(vocal)+0.5 L1(accomp.).	↑ SDR vocals 0.5 dB	1
C	**Phase-aware loss** reduces “hollow” artefacts.	Add `L_phase = 1 − cos Δφ` on 50 % randomly-selected bins; total loss = L_mag + 0.1 L_phase.	↑ SAR (fewer artefacts)	0.3
D	**Data augmentation** improves generalisation.	On-the-fly pitch-shift (±2 semitones) *or* time-stretch (0.9–1.1×) on mixture *and* stems before STFT.	↑ SDR 0.3 dB, cheap	0.2
E	**Curriculum on clip length** helps convergence.	Start with 3-s crops for 5 epochs → 6-s crops.	Faster convergence, same SDR	0
F	**Fine-tune Open-Unmix pre-trained** instead of training from scratch.	Load `umxhq` weights, unfreeze last GRU layer only, lr = 1×10^{-5}, 10 epochs.	↑ SDR 0.8 dB in 30 min	0.5

All runs: mixed precision, early stopping, save best-val ckpt.

3 Evaluation protocol (use for *every* run)

Level	Metric & tool
Track	SDR, SI-SDR, SAR with **`torchmetrics.audio`** (mirrors BSS-Eval v4).
Album (fold)	Report **median** over 50 val tracks; show iqr bars.
Stat. test	Wilcoxon signed-rank (paired) vs. baseline; $p < 0.01 \Rightarrow$ significant.
Qualitative	Waveform + mel-spectrogram side-by-side for 1 easy and 1 hard song.

4 One-week Colab schedule (≈ 8 GPU-h)

Day	What to do
1	Pipeline fixes; run baseline Open-Unmix-small 20 epochs.
2	Fine-tune full **umxhq** weights (Exp F).
3	Implement data aug & phase loss (Exp C + D); quick 15-epoch run.
4	Train **Demucs-tiny** (Exp A).
5	Multi-task mask head (Exp B) if time; else rerun best config with 3-fold CV.
6	Compute metrics, Wilcoxon p-values; render spectrogram figures.
7	Write Milestone 2: scoreboard, qualitative figs, compute budget, lessons.

(If Colab throttles GPU time, prioritise Exp F + Exp C + Exp D — all three finish in < 2 GPU-h.)

5 Colab survival tips

- **Cache** MUSDB18-HQ WAVs to Drive; down-mix to 32 kHz to halve I/O if quality OK.
- Use **chunked HDF5** for STFT magnitude tensors; skip re-computing each epoch.
- `TORCH_HOME=/content/drive/MyDrive/.cache/torch` keeps pre-trained weights across sessions.
- For Demucs time-domain models, cap **segment length 6 s** to avoid 12 h timeout.

6 Scoreboard template for the report

Exp	Backbone	Domain	Extras	SDR voc ↑	SI-SDR ↑	GPU min	Sig.?
Base	Open-Unmix-small	STFT	–	4.8	5.1	35	–
F	umxhq fine-tune	STFT	pre-trained	**5.6**	**6.0**	45	✓
C + D	Open-Unmix-small	STFT	phase + aug	5.4	5.8	40	✓
A	Demucs-tiny	time	–	5.1	**6.2**	80	✓

Shade best column values; ✓ when Wilcoxon p < 0.01 vs. base.

Quick-wins checklist

- Switch baseline to **STFT mask** (phase via mixture).
- **Fine-tune pre-trained Open-Unmix** (fastest gain).
- Add **phase-aware term** + **pitch/time aug** — almost free.
- Evaluate with **torchmetrics SDR / SI-SDR**, Wilcoxon stats.
- Log compute time & VRAM so choices are easy to justify.

Follow this compact plan to get a clear story about what moves the needle for vocal separation on limited compute. Good luck!