

Architettura Integrata per la Sintesi Musicale Neurale On-Premise: Progettazione, Implementazione e Ottimizzazione per l'Alta Fedeltà (2025)

1. Introduzione ed Executive Summary

Nel corso dell'ultimo biennio, il dominio della *Generative AI* applicata all'audio ha subito una metamorfosi strutturale, passando dalla semplice generazione di loop strumentali a bassa fedeltà verso la sintesi completa di brani musicali strutturati ("full-song generation"). La richiesta di un'infrastruttura *on-premise* capace di generare canzoni complete (testo, melodia, armonia, voce) con una "qualità estrema" non è più un desiderio futuristico, ma una possibilità concreta realizzabile attraverso l'orchestrazione di modelli *State-of-the-Art* (SOTA) rilasciati tra la fine del 2024 e l'inizio del 2025.

Questa relazione tecnica delinea una strategia esaustiva per la realizzazione di tale infrastruttura. L'obiettivo è superare i limiti qualitativi delle piattaforme cloud commerciali (come Suno o Udio) attraverso il controllo granulare offerto dall'esecuzione locale. La "qualità estrema" viene qui definita non solo in termini di specifiche tecniche (campionamento a 44.1kHz/48kHz, profondità di bit), ma come la convergenza di tre fattori critici: **coerenza strutturale** (la capacità del modello di mantenere una progressione logica tra strofe e ritornelli), **fedeltà timbrica** (la ricchezza spettrale degli strumenti e della voce) e **intelligibilità semantica** (la chiarezza e l'allineamento del testo cantato).

L'analisi identifica nel modello **YuE** (in particolare la variante YuE-s1-7B accoppiata a YuE-s2-1B) il nucleo generativo primario, grazie alla sua innovativa architettura ibrida che applica i paradigmi dei Large Language Models (LLM) alla sintesi audio.¹ Tuttavia, per raggiungere gli standard di eccellenza richiesti, un singolo modello non è sufficiente. L'architettura proposta si configura come una pipeline complessa orchestrata via **ComfyUI**, che integra LLM per la scrittura creativa (come **DeepSeek-V3** o **Llama 3 - GTP-oss-20b** finetuned), sistemi di separazione delle sorgenti (Stem Separation) basati su **UVR5**, motori di conversione vocale (Voice Conversion) tramite **RVC**, e suite di mastering automatizzato come

Matchering.³

Il presente documento offre una disamina dettagliata dei requisiti hardware (con focus specifico sulla gestione della VRAM per evitare la quantizzazione distruttiva), delle configurazioni software e dei flussi di lavoro di post-produzione necessari per mitigare gli artefatti tipici della sintesi neurale.

2. Fondamenti Teorici della Generazione Musicale Neurale (2025)

Per comprendere le scelte architettoniche proposte, è necessario analizzare l'evoluzione dei paradigmi di sintesi. Fino al 2023, la generazione musicale era dominata da due approcci distinti: i modelli di diffusione operanti su spettrogrammi (es. Riffusion, Stable Audio) e i modelli autoregressivi operanti su token audio discreti (es. MusicGen). Nel 2025, la frontiera si è spostata verso modelli unificati multimodali.

2.1. Il Paradigma "Lyrics-to-Audio" e la Tokenizzazione Semantica

La sfida storica nella generazione di canzoni complete risiedeva nell'allineamento temporale tra testo e musica. I modelli precedenti trattavano il testo come un condizionamento globale ("una canzone rock triste"), perdendo la sincronia sillabica. Il modello **YuE** (Multimedia Art Projection) ha risolto questo problema introducendo un vocabolario di token misto.

L'architettura di YuE non "vede" l'audio come un'onda continua, ma come una sequenza di codici discreti generati da un tokenizer neurale (simile a EnCodec o SoundStream). La novità risiede nell'interleaving (interlacciamento) di token testuali e token audio. Quando il modello genera, predice non solo il prossimo suono, ma anche la sua relazione con la sillaba corrente del testo. Questo approccio, definito "Music Language Modeling", permette di generare brani di diversi minuti mantenendo una coerenza narrativa e musicale impossibile per i modelli a diffusione pura.²

2.2. Architetture a Due Stadi: Disaccoppiamento di Struttura e Texture

Per raggiungere una "qualità estrema", è fondamentale separare la composizione dalla produzione sonora. I modelli *end-to-end* a singolo stadio spesso devono scendere a compromessi, sacrificando la complessità strutturale per la qualità audio o viceversa.

- **Stage 1 (Modellazione Semantica):** In questa fase, modelli massivi (come YuE-s1-7B) operano in uno spazio latente altamente compresso. L'obiettivo non è generare un suono piacevole, ma una struttura corretta: melodia, armonia, ritmo e prosodia. L'audio risultante da questo stadio è tecnicamente povero, ma musicalmente intelligente.
- **Stage 2 (Rifinitura Acustica):** Un secondo modello (es. YuE-s2-1B) agisce come un "upsampler" condizionato. Prende i token strutturali grezzi e "allucina" i dettagli ad alta frequenza, la risonanza degli strumenti e il timbro della voce. Questo disaccoppiamento permette di iterare sulla qualità sonora (rigenerando lo Stage 2) senza perdere la composizione (Stage 1).¹

2.3. L'Emergere di SongBloom e i Modelli Ibridi

Parallelamente a YuE, **SongBloom** introduce un'architettura che fonde la generazione autoregressiva (per lo "schizzo" musicale) con la diffusione (per la rifinitura). I modelli di diffusione sono intrinsecamente superiori nella gestione della fase e nella ricostruzione dei transienti (i picchi di energia iniziale dei suoni percussivi). Sebbene YuE sia superiore nella coerenza a lungo termine, SongBloom offre capacità di editing e *inpainting* (modifica di sezioni interne al brano) che lo rendono uno strumento complementare prezioso in una pipeline di produzione avanzata.⁷

3. Il Nucleo Generativo: Analisi Comparata e Selezione dei Modelli

La scelta del modello determina il tetto massimo della qualità raggiungibile. Analizziamo le opzioni disponibili per l'infrastruttura on-premise nel 2025.

3.1. YuE (YuE-s1-7B + YuE-s2-1B)

Attualmente, YuE rappresenta lo standard di riferimento per la generazione *open-weights* di

canzoni complete.

- **Punti di Forza:** Capacità di generare brani completi (strofa, ritornello, bridge) con una coerenza strutturale ineguagliata. Supporto nativo per il *In-Context Learning* (ICL), che permette di clonare lo stile o la voce da un file audio di riferimento di 30 secondi.⁹
- **Limitazioni:** Richiede risorse VRAM ingenti. La generazione è lenta rispetto ai modelli strumentali puri. La qualità del mixaggio stereo diretto può risultare talvolta "piatta" o con una separazione degli strumenti non ottimale rispetto a un mixaggio umano.
- **Verdetto per Qualità Estrema:** Indispensabile come motore primario per la composizione e la generazione vocale base.

3.2. SongBloom

- **Punti di Forza:** Eccellente nella gestione delle alte frequenze grazie al decoder a diffusione. Capacità di estendere brani esistenti o modificare parti specifiche senza rigenerare tutto il brano.
- **Limitazioni:** La coerenza del testo su brani molto lunghi può essere inferiore a YuE-7B. Minore supporto della community rispetto all'ecosistema YuE/ComfyUI attuale.¹⁰
- **Verdetto:** Utile come motore secondario per varianti o per generare sezioni strumentali complesse.

3.3. Stable Audio Open & MusicGen (Meta)

Questi modelli rimangono rilevanti per compiti specifici.

- **Stable Audio Open:** Ideale per generare effetti sonori (SFX), texture ambientali o intro strumentali cinemaniache dove la struttura canzone è meno rilevante della texture sonora.¹¹
- **MusicGen-Stem:** Fondamentale per la generazione multi-traccia nativa. Se l'obiettivo è avere il controllo totale sul mixaggio, MusicGen-Stem può generare basso, batteria e melodia su canali separati, permettendo un mixing professionale in post-produzione.¹³

3.4. Tabella Comparativa delle Capacità

Caratteristica	YuE-s1-7B	SongBloom	MusicGen Large	Stable Audio Open
Architettura	Autoregressive (LLM)	Hybrid (AR + Diffusion)	Autoregressive	Diffusion (DiT)
Generazione Vocale	Eccellente (Lyrics-aligned)	Buona	N/A (Solo Melodia)	N/A
Coerenza Strutturale	Molto Alta (>3 min)	Alta	Media	Bassa (<47 sec)
Fedeltà Audio	Alta (44.1kHz)	Alta	Media (32kHz)	Alta (44.1kHz)
Supporto ICL/Cloning	Sì (Zero-shot)	Limitato	No	No
Requisiti VRAM (FP16)	~24 GB	~16 GB	~16 GB	~8 GB

4. Architettura Hardware e Requisiti di Sistema

Per garantire la "qualità estrema", l'hardware non deve imporre compromessi software (come la quantizzazione aggressiva). L'analisi seguente dimensiona l'infrastruttura per un funzionamento ottimale.

4.1. Il Collo di Bottiglia della VRAM

I modelli audio basati su Transformer (come YuE) sono estremamente sensibili alla larghezza di banda della memoria e alla capacità totale della VRAM.

- **YuE-s1-7B (Stage 1):** In precisione mezza (FP16), il modello occupa circa 14-15 GB di VRAM. A questo si deve aggiungere lo spazio per il contesto (la "memoria" del brano generato finora) e i buffer CUDA. Una scheda da 16GB è il minimo teorico, ma rischia

errori *Out-Of-Memory* (OOM) su brani lunghi.

- **Qualità vs. Quantizzazione:** È tecnicamente possibile eseguire questi modelli su schede da 8GB o 12GB usando la quantizzazione a 8-bit o 4-bit (es. GGUF/EXL2). Tuttavia, nel dominio audio, la quantizzazione degrada percettibilmente la qualità: i transienti perdono impatto, le code dei riverberi diventano metalliche e la stabilità della voce peggiora.¹⁵ Per l'obiettivo "qualità estrema", l'esecuzione in **FP16** o **BF16** è mandataria.

4.2. Configurazione Hardware Raccomandata — NON NECESSARIA

1. **GPU (Graphics Processing Unit):**
 - **Gold Standard: NVIDIA RTX 4090 (24GB).** È la scelta ottimale per il rapporto prezzo/prestazioni. I 24GB permettono di caricare YuE-7B in FP16 lasciando spazio per generazioni lunghe.
 - **Workstation Tier: NVIDIA RTX 6000 Ada (48GB)** o configurazioni multi-GPU (**2x RTX 3090/4090**). Due GPU da 24GB permettono di tenere caricati contemporaneamente il modello di generazione (YuE) e i modelli di post-produzione (RVC, UVR5), accelerando drasticamente il workflow iterativo.¹⁶
2. **CPU:** Un processore moderno con un alto numero di core (es. AMD Ryzen 9 7950X o Intel i9-14900K) è necessario per il preprocessing dei dati audio e per gestire i modelli di separazione (UVR5) che beneficiano anche della parallelizzazione CPU se la GPU è occupata.
3. **RAM di Sistema: 64 GB DDR5 (minimo) o 128 GB.** Quando si passa da un modello all'altro (es. da generazione a separazione stem), i pesi vengono spostati dalla VRAM alla RAM. Una RAM capiente e veloce minimizza i tempi morti.
4. **Storage: SSD NVMe Gen4 (o Gen5)** da almeno 2TB. I modelli audio non compressi e i dataset di inferenza occupano spazio significativo. La velocità di lettura è critica per il caricamento dei modelli.

4.3. Ottimizzazione Software (CUDA & Linux)

Sebbene Windows sia supportato, un ambiente **Linux (Ubuntu 22.04/24.04)** è fortemente consigliato per l'infrastruttura di produzione. Linux offre una gestione della memoria VRAM più efficiente (senza l'overhead del WDDM di Windows) e supporta nativamente tecnologie come **Flash Attention 2**, cruciali per velocizzare l'inferenza dei Transformer audio lunghi.² Se si è vincolati a Windows, l'uso di **WSL2** (Windows Subsystem for Linux) è un compromesso accettabile.

5. Implementazione del Workflow: L'Ecosistema ComfyUI

L'orchestrazione di modelli eterogenei richiede un framework flessibile. **ComfyUI**, con la sua architettura a nodi, si è affermato come lo standard industriale per la generazione AI locale.

5.1. Configurazione dell'Ambiente ComfyUI

Per abilitare la generazione musicale, è necessario estendere l'installazione base di ComfyUI con nodi custom specifici:

- **ComfyUI_YuE:** Il wrapper essenziale per caricare ed eseguire i modelli YuE. Gestisce sia lo Stage 1 che lo Stage 2 e l'integrazione ICL.¹⁸
- **ComfyUI-Audio:** Fornisce le primitive per caricare, salvare, visualizzare e manipolare le forme d'onda audio all'interno del grafo.
- **ComfyUI_RVC:** Permette di eseguire l'inferenza RVC (Voice Conversion) direttamente nel workflow, senza dover esportare l'audio e processarlo esternamente.
- **ComfyUI-Essentials:** Una suite di utility per la gestione logica del flusso.

5.2. Pipeline Operativa Dettagliata (Step-by-Step)

Il workflow per la "qualità estrema" si articola in cinque fasi sequenziali, implementate come gruppi di nodi in ComfyUI.

Fase 1: Ideazione e Strutturazione Semantica (Lyrics Generation)

La qualità della canzone inizia dal testo. I modelli audio sono sensibili alla metrica e alla struttura.

- **Nodo LLM:** Si integra un modello di linguaggio locale (es. **DeepSeek-V3** o **Llama 3-8B**

Instruct finetuned per scrittura creativa).³

- **Prompt Engineering:** Il prompt al LLM deve richiedere esplicitamente la struttura con tag che YuE riconosce:
 - [Verse], [Chorus], `` , [Outro].
 - Istruzioni di stile nel prompt musicale: "Genre: Cinematic Pop, Female Vocals, Slow Tempo, Reverb-heavy piano".

Fase 2: Generazione Core (YuE Stage 1 & 2)

- **Configurazione YuE:**
 - *Input:* Testo strutturato dal nodo LLM + Prompt di Genere.
 - *ICL (Opzionale ma Raccomandato):* Per la massima qualità, si fornisce al nodo un file audio di riferimento (15-30 secondi) che possiede le caratteristiche timbriche desiderate. Questo guida lo Stage 1 verso uno stile di produzione professionale.⁹
 - *Parametri:* Temperature a 1.0 per creatività, Top-K a 50. È cruciale impostare il Repetition Penalty a 1.1 o 1.2 per evitare loop infiniti nel finale dei brani.
- **Output Intermedio:** Un file audio grezzo che contiene il mix completo (voce + musica).

Fase 3: Separazione degli Stem (UVR5)

L'audio generato da YuE, sebbene coerente, è un "flat mix". Per raggiungere la qualità estrema, dobbiamo separare gli elementi per processarli individualmente.

- **Nodo UVR5:** Il file audio viene passato a un nodo che esegue **Ultimate Vocal Remover**.
- **Selezione Modello:**
 - Per la Voce: Kim_Vocal_2 o MDX23C-InstVoc HQ. Questi modelli sono SOTA per estrarre voci pulite riducendo al minimo il "bleeding" degli strumenti.²¹
 - Per la Strumentale: UVR-MDX-NET Inst HQ 3 o Demucs v4 (Hybrid Transformer).
- **Risultato:** Due tracce audio separate: Vocals.wav e Instrumental.wav.

Fase 4: Rifinitura Vocale (Voice Conversion - RVC)

La voce generata da YuE può talvolta suonare leggermente "granulosa" o artificiale.

- **Processo RVC:** La traccia Vocals.wav viene processata da un modello RVC.

- **Strategia di Qualità:** Si utilizza un modello RVC addestrato su una voce umana di altissima qualità (studio recording). L'obiettivo non è necessariamente cambiare l'identità del cantante, ma trasferire la *texture* ad alta definizione della voce umana sulla performance intonata generata da YuE.
- **Impostazioni:** Indice di feature (Index Rate) basso (0.3 - 0.5) per mantenere l'espressività originale di YuE ma migliorare il timbro.²³

Fase 5: Mastering Automatizzato (Matchering)

L'ultimo passaggio è ricomporre il brano e dargli il "suono" di un prodotto commerciale.

- **Nodo Mixing:** Ricombina Refined_Vocals.wav e Instrumental.wav. Si possono applicare leggeri EQ o riverberi in questa fase.
 - **Nodo Matchering:** Utilizza la libreria **Matchering 2.0**.⁵
 - *Input:* Il mix grezzo.
 - *Reference:* Un brano commerciale di riferimento (caricato dall'utente).
 - *Processo:* L'algoritmo analizza la curva spettrale, l'RMS e l'immagine stereo del riferimento e la applica al target, garantendo un bilanciamento professionale e un volume competitivo (-14 LUFS).
-

6. Strategie Avanzate per l'Alta Fedeltà (Deep Dive)

Oltre al workflow standard, esistono tecniche avanzate per spingere la qualità al limite fisico dei modelli attuali.

6.1. Audio Super-Resolution e Bandwidth Extension

Molti modelli audio (incluso YuE nelle configurazioni standard) possono avere un *roll-off* (taglio) delle frequenze sopra i 16kHz o 20kHz, risultando in un suono meno "ariosa" rispetto a una registrazione a 44.1kHz/96kHz reale.

- **Soluzione:** Integrare nel workflow nodi di **Audio Super-Resolution** (come **AudioUNet** o modelli basati su GAN). Questi modelli "immaginano" e ricostruiscono le frequenze ultra-alte basandosi sulle armoniche presenti nelle frequenze medie, restituendo

brillantezza ai piatti della batteria e alle sibilanti vocali.²⁴

6.2. Gestione della Fase e Allineamento Temporale

Quando si separano e si ricombinano le tracce (workflow UVR -> RVC -> Mix), si rischia di introdurre problemi di fase che cancellano alcune frequenze (suono "scatolato").

- **Mitigazione:** È fondamentale che l'intera catena di elaborazione operi con precisione al campione. In ComfyUI, assicurarsi che non ci siano nodi che introducono latenza non compensata (es. alcuni effetti VST wrapper). L'uso di RVC in modalità "streaming" o con chunking aggressivo può introdurre disallineamenti; è preferibile processare l'intero file in una singola passata (batch processing) quando la VRAM lo consente.

6.3. In-Context Learning (ICL) per il Controllo Stilistico

L'ICL è la chiave per superare la genericità del suono AI. Invece di affidarsi solo al prompt testuale ("Rock song"), fornire a YuE due prompt audio:

1. **Vocal Prompt:** 10 secondi di una voce isolata con il timbro desiderato.
2. Instrumental Prompt: 10 secondi di una base con il sound design desiderato.

YuE è in grado di fondere questi due input, generando una nuova composizione che mantiene la "pasta sonora" degli input. Questo è particolarmente potente per creare interi album con un suono coerente.⁹

7. Implementazione Pratica: Installazione e Codice

Questa sezione fornisce dettagli operativi per l'installazione dei componenti core.

7.1. Preparazione dell'Ambiente Python

Si consiglia l'uso di conda per isolare le dipendenze.

Bash

```
# Creazione environment
conda create -n ai-music python=3.10
conda activate ai-music

# Installazione PyTorch con supporto CUDA 12.1 (Verificare driver)
pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu121

# Installazione ComfyUI
git clone https://github.com/comfyanonymous/ComfyUI
cd ComfyUI
pip install -r requirements.txt
```

7.2. Installazione dei Modelli YuE

Il download dei pesi richiede spazio e banda. Si consiglia di scaricare i modelli anneal-en-cot (Chain of Thought, Inglese/Multilingua) per la migliore coerenza.

Bash

```
# Navigare nella cartella modelli di ComfyUI
cd models/checkpoints
# Creare cartella dedicata
mkdir yue_models
cd yue_models

# Esempio di download (richiede huggingface-cli)
huggingface-cli download m-a-p/YuE-s1-7B-anneal-en-cot --local-dir YuE-s1-7B
huggingface-cli download m-a-p/YuE-s2-1B-general --local-dir YuE-s2-1B
```

Nota: Assicurarsi di scaricare anche i file di configurazione (config.json, tokenizer.model) e non solo i tensori .safetensors.¹⁹

7.3. Integrazione di Matchering

Poiché Matchering è una libreria Python e non un nodo nativo standard, potrebbe essere necessario installarlo nell'ambiente di ComfyUI e utilizzare un nodo script personalizzato o un wrapper esterno.

Bash

```
pip install matchering
```

In uno script Python o nodo custom:

Python

```
import matchering as mg
# Processo semplice
mg.process(
    target="generated_song.wav",
    reference="reference_master.wav",
    results=[
        mg.pcm16("mastered_song.wav"),
    ],
)
```

8. Considerazioni sulle Limitazioni e Risoluzione Problemi

8.1. Allucinazioni Musicali

- **Problema:** Il modello genera audio senza senso o rumore statico dopo il secondo minuto.
- **Causa:** Accumulo di errori nel contesto autoregressivo o repetition_penalty troppo basso.
- **Soluzione:** Aumentare il repetition_penalty a 1.2. Provare a generare in segmenti più corti e unirli (inpainting/continuation) se il modello supporta la gestione dello stato.²

8.2. Qualità del Testo Italiano

- **Problema:** La pronuncia italiana risulta "inglesizzata".
- **Causa:** Il dataset di training è prevalentemente inglese/cinese.
- **Soluzione:** Utilizzare la fonetizzazione esplicita nel prompt testuale (scrivere come si pronuncia) o, per risultati professionali, utilizzare il workflow RVC: generare la base con una voce "guida" qualsiasi, e poi sostituirla con un modello RVC addestrato specificamente su un madrelingua italiano.

9. Conclusioni

L'infrastruttura descritta rappresenta il vertice tecnologico accessibile nel 2025 per la produzione musicale AI on-premise. L'integrazione di **YuE** per la struttura compositiva, **UVR5** per la chirurgia spettrale, **RVC** per la texture vocale e **Matchering** per la finalizzazione dinamica permette di superare i limiti qualitativi dei singoli modelli "monolitici".

Sebbene l'investimento hardware (GPU da 24GB+) e la curva di apprendimento (ComfyUI) siano significativi, il risultato è una *pipeline* di produzione sovrana, priva di costi di licenza ricorrenti e capace di generare output di fedeltà "estrema" indistinguibili, ad un orecchio non allenato, da produzioni commerciali umane. Il futuro di questa tecnologia risiede ora nel perfezionamento dei modelli di linguaggio per una migliore aderenza alle sfumature emotive e culturali delle diverse lingue, ma l'architettura qui definita è pronta per accogliere queste evoluzioni modulari.

Bibliografia

1. m-a-p/YuE-s1-7B-anneal-en-icl - Hugging Face, accesso eseguito il giorno dicembre 1, 2025, <https://huggingface.co/m-a-p/YuE-s1-7B-anneal-en-icl>

2. YuE: AI Music Generation at Scale - WhiteFiber, accesso eseguito il giorno dicembre 1, 2025, <https://www.whitefiber.com/blog/yue-ai-music-generator>
3. The Best Open Source LLM For Creative Writing & Ideation In 2025 - SiliconFlow, accesso eseguito il giorno dicembre 1, 2025, <https://www.siliconflow.com/articles/en/best-open-source-lm-for-creative-writing-ideation>
4. Best AI Vocal Removers in 2025 - DEV Community, accesso eseguito il giorno dicembre 1, 2025, <https://dev.to/lalalai/best-ai-vocal-removers-in-2025-1036>
5. Matchering 2.0 - LinuxMusicians, accesso eseguito il giorno dicembre 1, 2025, <https://linuxmusicians.com/viewtopic.php?t=21053>
6. YuE: Scaling Open Foundation Models for Long-Form Music Generation - arXiv, accesso eseguito il giorno dicembre 1, 2025, <https://arxiv.org/html/2503.08638v1>
7. YuE: Scaling Open Foundation Models for Long-Form Music Generation | OpenReview, accesso eseguito il giorno dicembre 1, 2025, <https://openreview.net/forum?id=hZy6YG2lj8>
8. SongBloom: Coherent Song Generation via Interleaved Autoregressive Sketching and Diffusion Refinement - Microsoft Research, accesso eseguito il giorno dicembre 1, 2025, <https://www.microsoft.com/en-us/research/publication/songbloom-coherent-song-generation-via-interleaved-autoregressive-sketching-and-diffusion-refinement/>
9. deepbeepmeep/YuEGP: YuE: Open Full-song Generation Foundation for the GPU Poor - GitHub, accesso eseguito il giorno dicembre 1, 2025, <https://github.com/deepbeepmeep/YuEGP>
10. Tencent SongBloom music generator updated model just dropped. Music + Lyrics, 4min songs. : r/StableDiffusion - Reddit, accesso eseguito il giorno dicembre 1, 2025, https://www.reddit.com/r/StableDiffusion/comments/1okpsj4/tencent_songbloom_music_generator_updated_model/
11. (PDF) Stable Audio Open - ResearchGate, accesso eseguito il giorno dicembre 1, 2025, https://www.researchgate.net/publication/382445394_Stable_Audio_Open
12. AI Solutions for Everywhere Your Sounds Shows Up | Stable Audio 2.5 - Stability AI, accesso eseguito il giorno dicembre 1, 2025, <https://stability.ai/stable-audio>
13. MusicGen-Stem: Multi-stem music generation and edition through autoregressive modeling, accesso eseguito il giorno dicembre 1, 2025, <https://arxiv.org/html/2501.01757v1>
14. facebook/musicgen-stem-6cb at main - Hugging Face, accesso eseguito il giorno dicembre 1, 2025, <https://huggingface.co/facebook/musicgen-stem-6cb/tree/main>
15. YuE GP, runs the best open source song generator with less than 10 GB of VRAM - Reddit, accesso eseguito il giorno dicembre 1, 2025, https://www.reddit.com/r/StableDiffusion/comments/1iegcx/yue_gp_runs_the_best_open_source_song_generator/
16. YuE - Local Music Generation with Audio Prompts - FOSS - 6GB VRAM! - YouTube, accesso eseguito il giorno dicembre 1, 2025,

<https://www.youtube.com/watch?v=6FBnKljqT04>

17. Fish TTS: Master Custom Voice Creation – Step-by-Step Installation Guide! - YouTube, accesso eseguito il giorno dicembre 1, 2025,
<https://www.youtube.com/watch?v=VwTiUoNsXTs>
18. ComfyUI_YuE detailed guide | ComfyUI - RunComfy, accesso eseguito il giorno dicembre 1, 2025, https://www.runcomfy.com/comfyui-nodes/ComfyUI_YuE
19. smthemex/ComfyUI_YuE: YuE is a groundbreaking series of open-source foundation models designed for music generation, specifically for transforming lyrics into full songs (lyrics2song). you can use it in comfyUI - GitHub, accesso eseguito il giorno dicembre 1, 2025, https://github.com/smthemex/ComfyUI_YuE
20. Llama 3 Fundamentals Full Course | Master LLMs, Fine-Tuning, Hugging Face and LoRA, accesso eseguito il giorno dicembre 1, 2025,
<https://www.youtube.com/watch?v=qF9WceD5JbY>
21. How to Use Ultimate Vocal Remover [2025 Newest Tips], accesso eseguito il giorno dicembre 1, 2025,
<https://vocalremover.easeus.com/ai-article/how-to-use-ultimate-vocal-remover.html>
22. What are the best models for clean vocal extract? · Anjok07 · ultimatevocalremovergui · Discussion #689 - GitHub, accesso eseguito il giorno dicembre 1, 2025,
<https://github.com/Anjok07/ultimatevocalremovergui/discussions/689>
23. Advanced Techniques for Post Processing AI Vocals Mixes : r/IsolatedTracks - Reddit, accesso eseguito il giorno dicembre 1, 2025,
https://www.reddit.com/r/IsolatedTracks/comments/1fd8bb7/advanced_techniques_for_post_processing_ai_vocals/
24. yuvraj108c/ComfyUI_InvSR: ComfyUI wrapper for InvSR (Arbitrary-steps Image Super-resolution via Diffusion Inversion) - GitHub, accesso eseguito il giorno dicembre 1, 2025, https://github.com/yuvraj108c/ComfyUI_InvSR