ACM Music

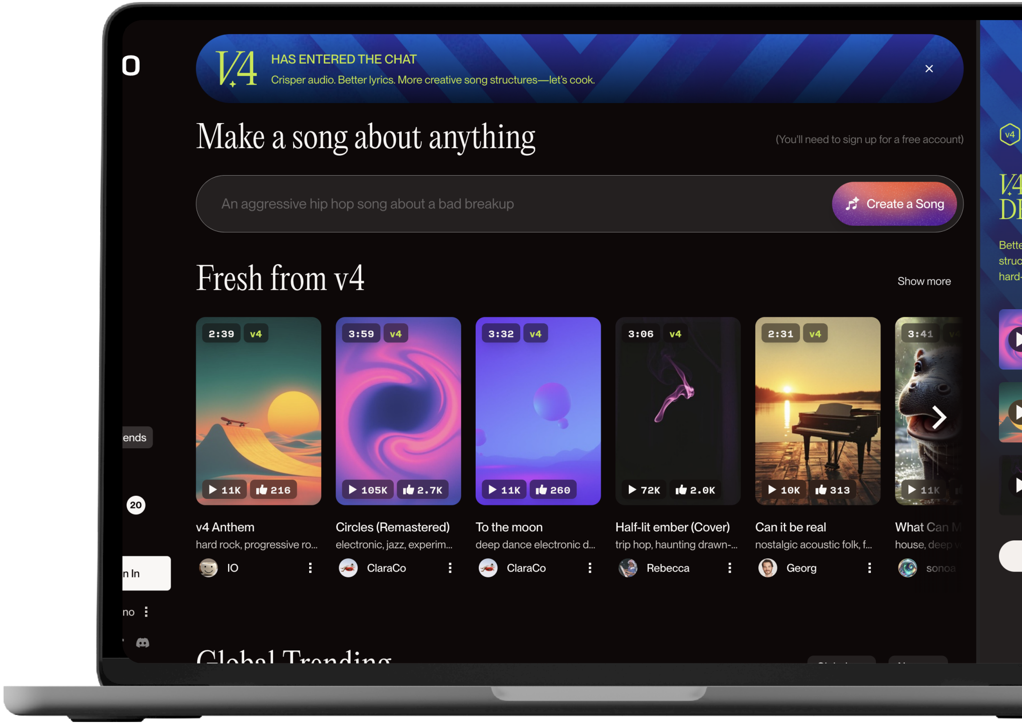
An Affordable, Controllable and Modular Music Generation System

Music Generation is Simple

Just Two Steps

1. Step #1: Write Prompt and Submit
2. Step #2: Download Generated Music

But How?



V4 HAS ENTERED THE CHAT

Crisper audio. Better lyrics. More creative song structures—let's cook.

Make a song about anything

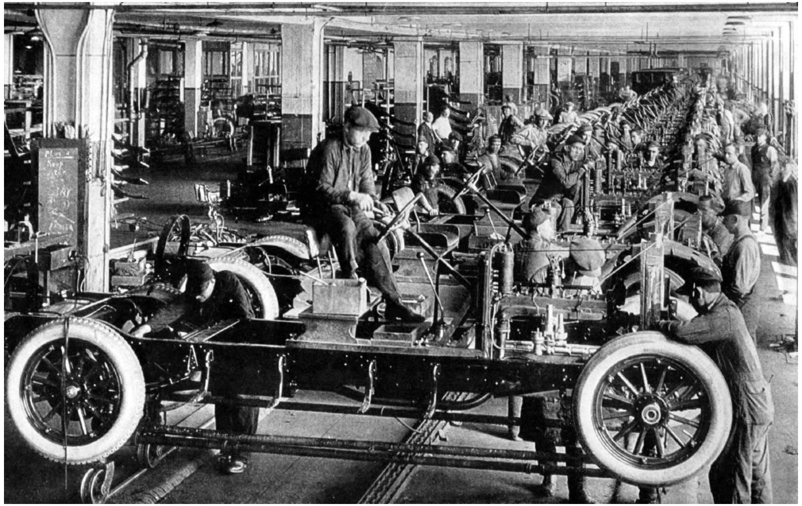
(You'll need to sign up for a free account)

An aggressive hip hop song about a bad breakup

Fresh from v4

Fordism: Step by Step or End-to-End

1930s:
A Master Craftsman building the entire car or
An Assembly Line of specialized workers
2020s:
A Large and Expensive Model or
A Pipeline comprised with Small Models



The design philosophy of our music generation system draws inspiration from Fordist principles of modular manufacturing, applying them to AI architecture.

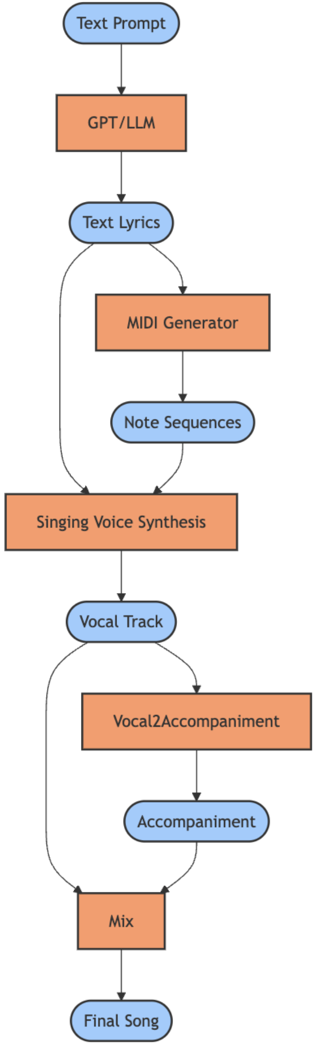
Modular Pipeline

01 The Text Prompt stage initiates the pipeline by accepting user input. This is where you describe your desired song concept, mood, style, theme, or story. The prompt should be clear and detailed enough to guide the AI in generating appropriate lyrics that match your creative vision.

02 The Lyric and Melody Generation phase transforms your text prompt into musical elements. The GPT/LLM model generates structured lyrics, while the MIDI Generator creates matching melodies. These two components work together to ensure the lyrics and musical notes are coherent and complementary, forming the foundation of your song.

03 The Voice Synthesis stage combines the generated lyrics and note sequences to create the vocal track. This critical phase aligns the text with the melody, ensuring proper timing and natural-sounding pronunciation. The system synthesizes a human-like singing voice that follows the musical composition while maintaining clarity of the lyrics.

04 The Final Production stage handles the musical arrangement and mixing. The Vocal2Accompaniment system generates appropriate instrumental accompaniment based on the vocal track. Finally, the Mix process combines the synthesized vocals with the generated accompaniment, balancing all elements to produce a complete, professionally-structured final song.



Interpretability and Control

- Each module has a clear, single responsibility (lyrics, melody, voice, accompaniment)
- Outputs at each stage can be monitored and adjusted
- Problems can be isolated to specific components
- Artists can intervene at any stage of the creative process

Flexibility and Customization

- Modules can be upgraded independently as technology improves
- Different models can be swapped in for each component
- Easy to adapt to different musical styles and requirements
- Components can be reused in other applications

Maintainability and Development

- Teams can specialize in improving specific modules
- Easier debugging and quality control
- Reduced complexity in training and optimization
- More manageable development cycles

Resource Efficiency

- Each module can be optimized independently
- No need to retrain the entire system for improvements
- Can distribute computation across different resources
- More efficient use of training data

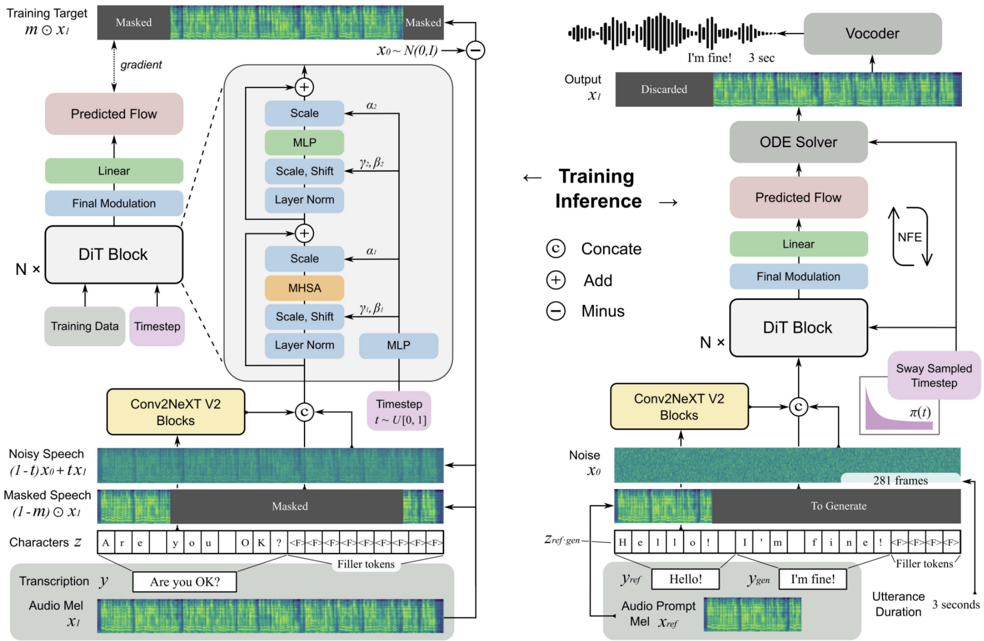
Modules	ACM Music	OpenSuno
Lyric Generation	**Just use LLM Services**	100 x H100 GPUs 20K Hours Recording
Lyric-to-Melody	**4 x A100** **2000 Pairs of lyric and melody data**	
Singing Voice Synthesis	**4 x A100** **About 10 hours singing data**	
Vocal2Accompaniment	**8 x A100** **5K Hours Recording**	

Resource Efficiency

- Each module can be optimized independently
- No need to retrain the entire system for improvements
- Can distribute computation across different resources
- More efficient use of training data

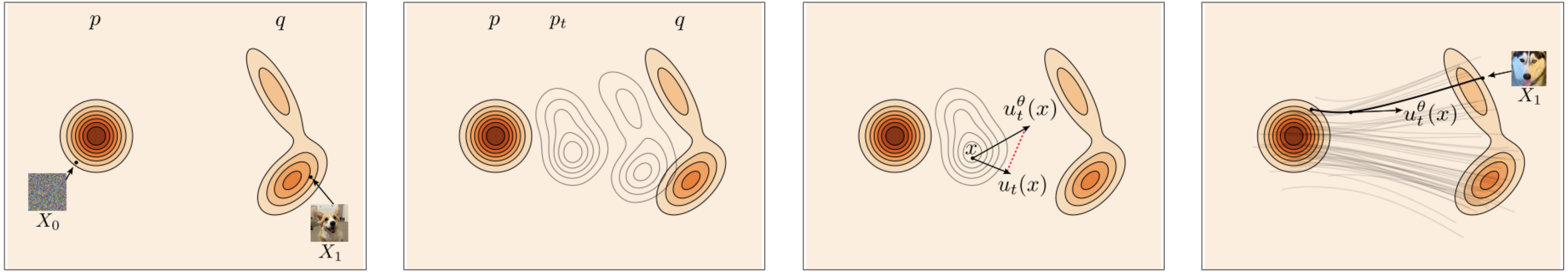
Modules	ACM Music	OpenSuno
Lyric Generation	**Just use LLM Services**	100 x H100 GPUs 20K Hours Recording
Lyric-to-Melody	**4 x A100** **2000 Pairs of lyric and melody data**	
Singing Voice Synthesis	**4 x A100** **About 10 hours singing data**	
Vocal2Accompaniment	**8 x A100** **5K Hours Recording**	

Singing Voice Sythesis



- Flowing Matching Based Model
- DiT for Predication of Velocity Field
- Conv2NeXT for Processing Text Input
- Masked Training Strategy for Audio Prompting
- ODE Solver: Euler or Mid Point
- Classifier-Free Guidance for Conditional Generation

Singing Voice Sythesis



(a) Data. **(b)** Path design. **(c)** Training. **(d)** Sampling.

Singing Voice Sythesis

SVS Training Data Curation

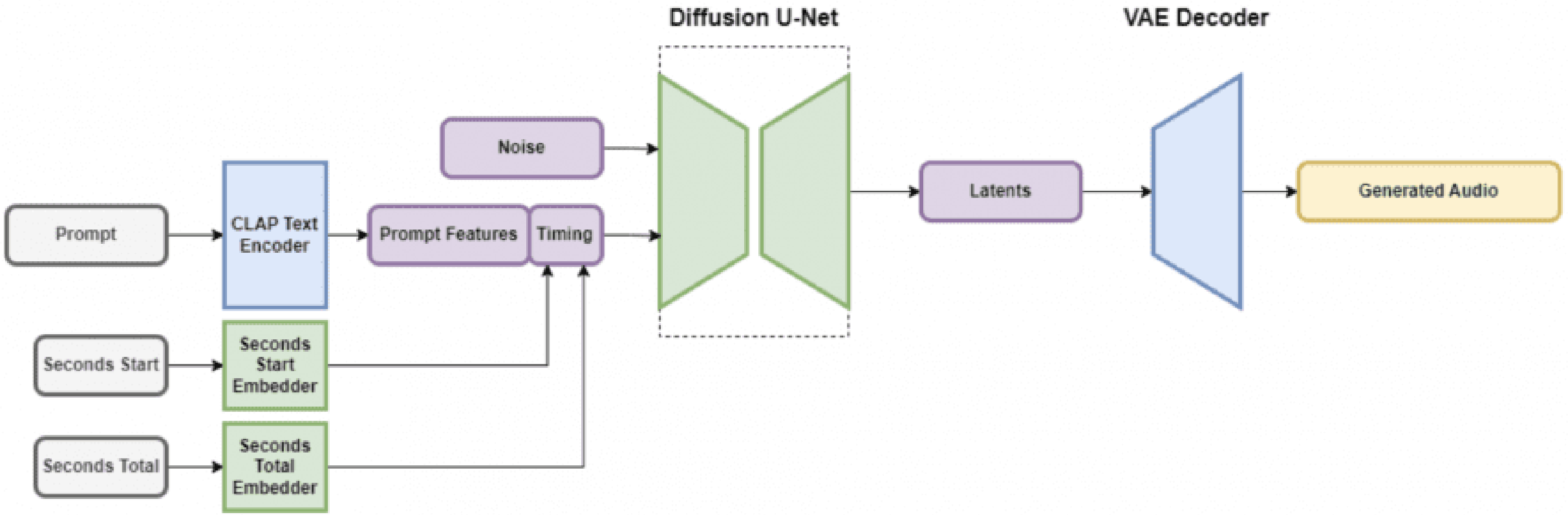
- M4Singer Dataset which contains 29.77 hours of high-quality Mandarin singing voice without accompaniment
- All recording are sliced in sentence levels
- For Simplicity of impl of dataset loading, we use special marks for indicate notes:
    - ⓪你❼流❽连❼❽电❼影❼❽里❼美⓿丽❸的❸不❸真❸实❺的❻❼场❺景
    - 'A#3': '⓪', 'B3': '⓿', 'C4': '❶', 'C#4': '❷', 'D4': '❸',
    - 'D#4': '❹', 'E4': '❺', 'F4': '❻', 'F#4': '❼', 'G4': '❽',
    - 'G#4': '❾', 'A4': '❿', 'A#4': '⓫', 'B4': '⓬',
- 8 x A100 40G x 36 Hours

SVS: Advantages and Drawbacks

- Few-Shots Speaker Cloning (5-10 seconds); Can clone accent
- No Need for explicitly indicate the duration (The output lacks good rhyme also)

- Hissing Noise when pitch go high
    - Finetune the vocoder of TTS for Singing
- Sing syllable by syllable
    - Finetune a language model for inducing duration information from lyrics
- Can't generate English Singing
    - Curate English Singing Data by using Voice Transcription Model
- Lack a set of Comphrehensive Evaluation Metrics for SVS

Singing2Accompaniment



- DiT Based S2A Model
- Finetuned from Stable-Audio-Open

S2A Data Curation and Training

- Scrapped Mandarin Song Dataset Contains 3000 recordings
- All recording are separated to Vocal and Instrumental Parts
- Use ensembled Source Separation for Better Quality
    - Band-Split RoFormer
    - Demucs
- 8 x A100 40G x 24 Hours

S2A Data Curation and Training

- Scrapped Mandarin Song Dataset Contains 3000 recordings
- All recording are separated to Vocal and Instrumental Parts
- Use ensembled Source Separation for Better Quality
  - Band-Split RoFormer
  - Demucs
- 8 x A100 40G x 24 Hours
- For allievate the issue of Info Leackage caused by Artifcats of Source Separation, the vocals are applied noise and distoration before fed into conidition embedding extractor

S2A Samples

	Origianl	Vocals	Origianl Vocals + Generated Acc
“Plastic Love”	🔊	🔊	🔊
“Even savages are capable of love“	🔊	🔊	🔊

S2A: Advantages and Drawbacks

- Training and Inference are much more faster than Token LM based methods like SingSong
- The generated accompaniment clashes with the vocals
- Solution: Also utilize the Chord Information for Prompting
    - Use Pretrained Chord Recognition Model
- Lack of Evaluation Metrics