PAT 498/598 (Fall 2024)

Special Topics:
Generative AI for Music and Audio Creation

Lecture 20: Interactive & Multimodal Systems

Instructor: Hao-Wen Dong

SCHOOL OF MUSIC, THEATRE & DANCE
PERFORMING ARTS TECHNOLOGY
UNIVERSITY OF MICHIGAN

Final Project

- Milestones (all due at the specified date at **11:59 PM ET**)

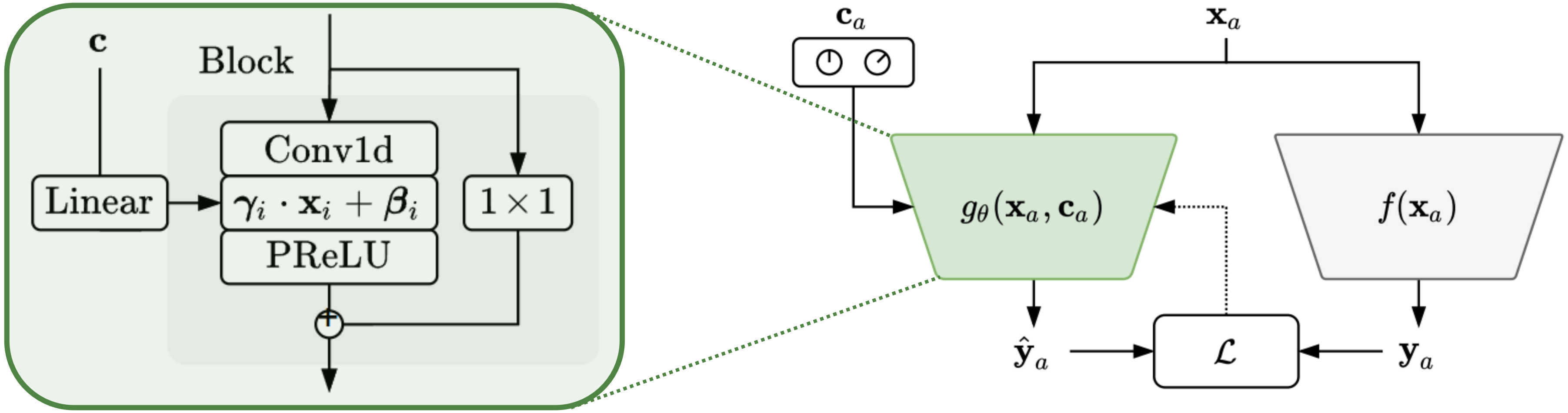
Pitch	November 6	Topic & high-level plans
Proposal	November 22	Survey & plans (1 page)
Presentation	December 9	Showcase & report
Final report	December 15	Full report (3-5 pages)

- Instructions will be released on Gradescope
- Late submissions: **NOT accepted**

Final Project Rubrics

Proposal	**10pt**	
Presentation	**20pt**	
Final report	**30pt**	
Implementation		10pt
Code documentation		5pt
Explanation of design and implementation		5pt
Results, analysis and discussions		10pt

(Recap) Example: Neural Audio Effects (Steinmetz et al., 2021)

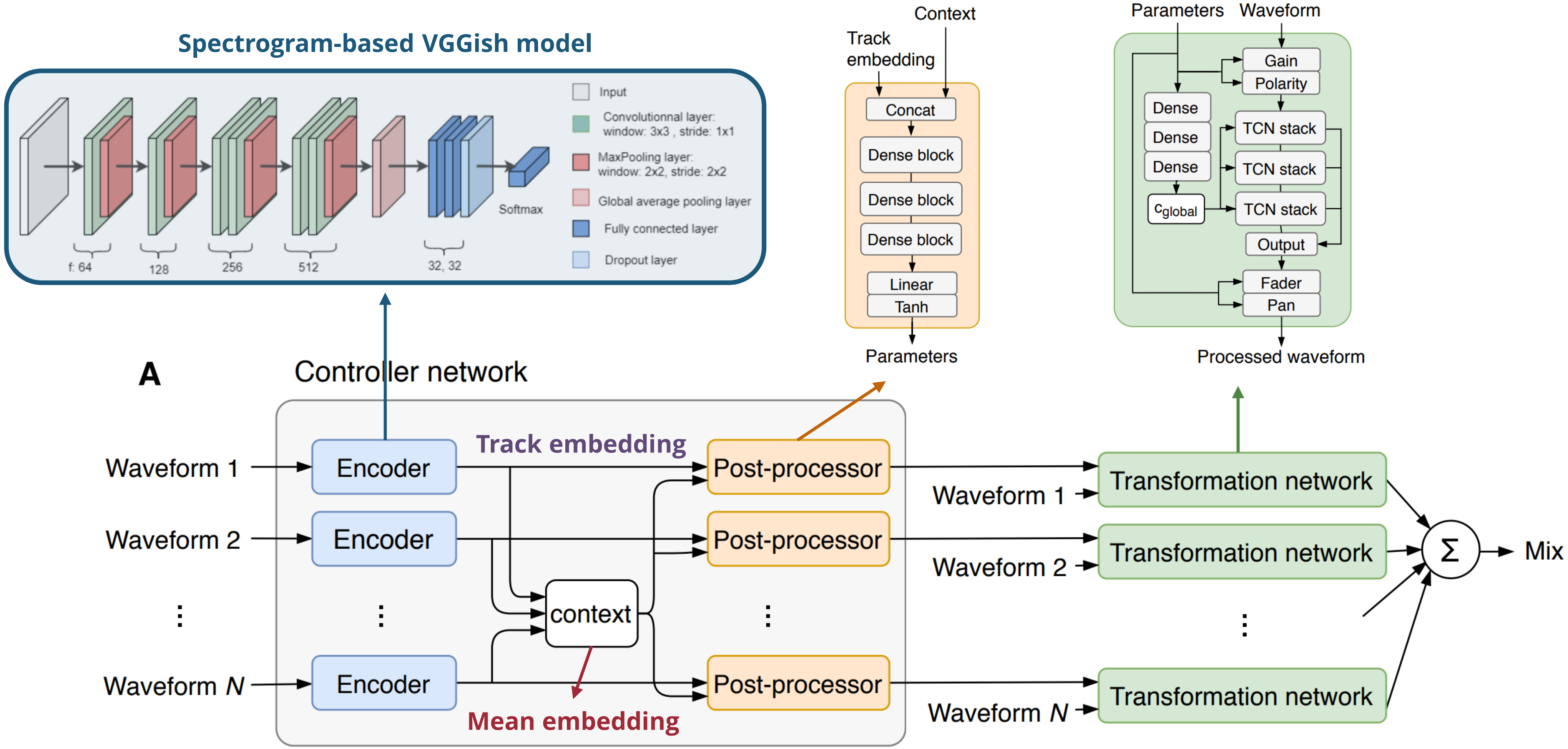


(Source: Steinmetz et al., 2021)

csteinmetz1.github.io/steerable-nafx

Christian J. Steinmetz and Joshua D. Reiss, "Steerable discovery of neural audio effects," *NeurIPS ML4CD Workshop*, 2021.

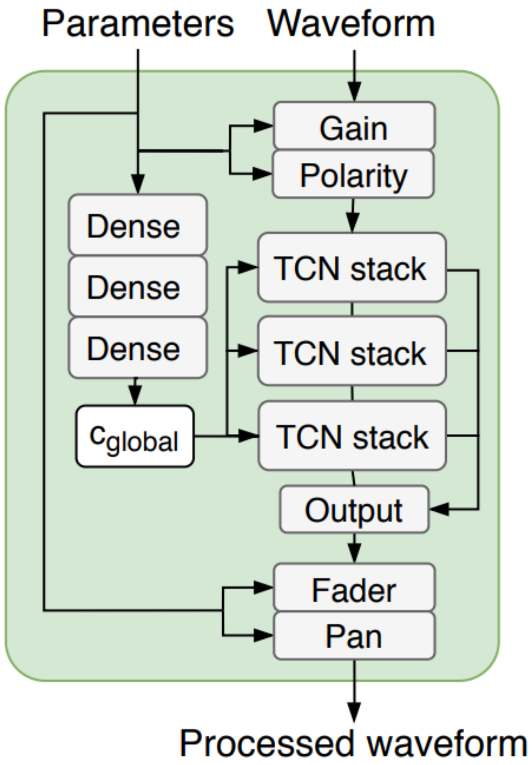
(Recap) Example: Differentiable Auto-mixing (Steinmetz et al., 2021)



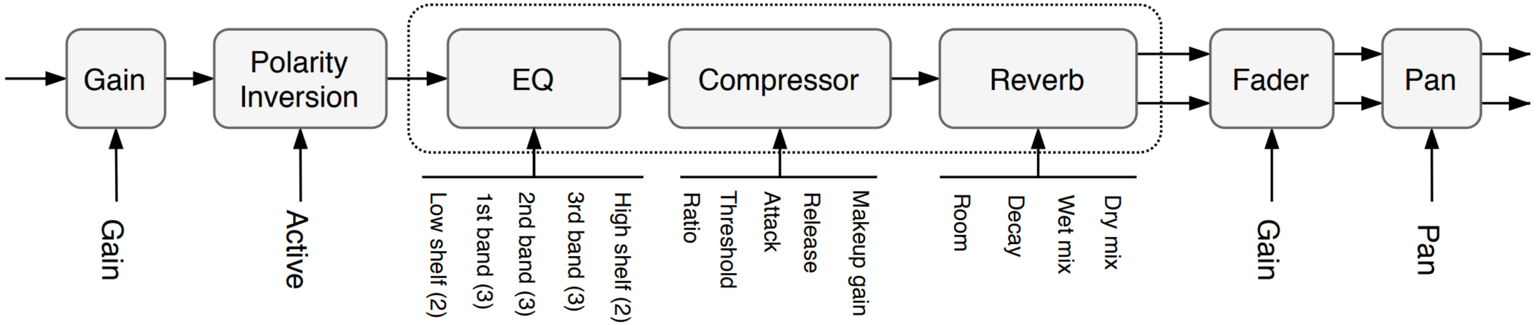
(Source: Steinmetz et al., 2021)

Christian J. Steinmetz, Jordi Pons, Santiago Pascual, and Joan Serrà, “Automatic multitrack mixing with a differentiable mixing console of neural audio effects,” *ICASSP*, 2021.

(Recap) Example: Differentiable Auto-mixing (Steinmetz et al., 2021)



(Source: Steinmetz et al., 2021)



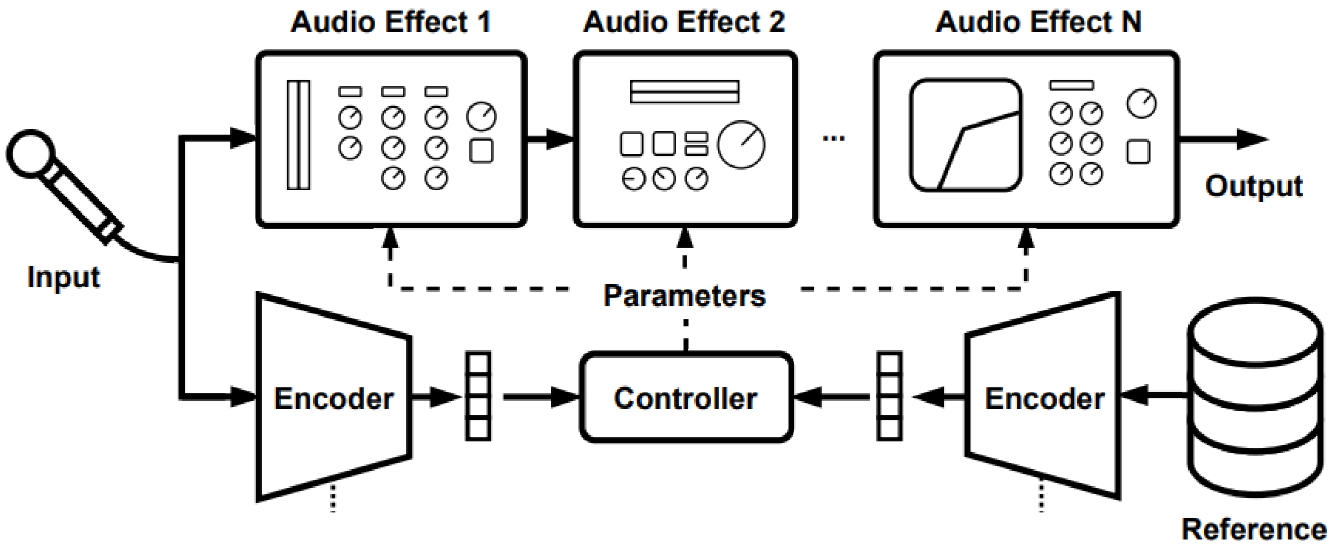
(Source: Steinmetz et al., 2021)

A differentiable (and thus trainable) mixing console!

github.com/csteinmetz1/pymixconsole

Christian J. Steinmetz, Jordi Pons, Santiago Pascual, and Joan Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," *ICASSP*, 2021.

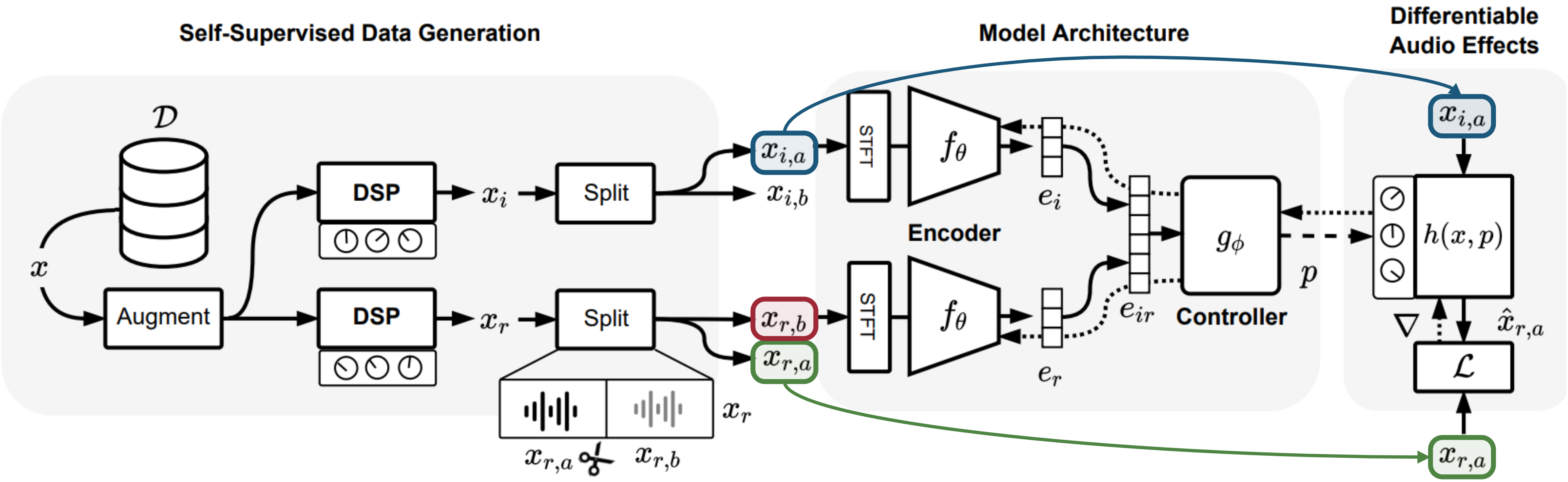
(Recap) Example: DeepAFx-ST (Steinmetz et al., 2022)



(Source: Steinmetz et al., 2022)

Christian J. Steinmetz, Nicholas J. Bryan, and Joshua D. Reiss, "Style Transfer of Audio Effects with Differentiable Signal Processing," *JAES*, 2022.

(Recap) Example: DeepAFx-ST (Steinmetz et al., 2022)

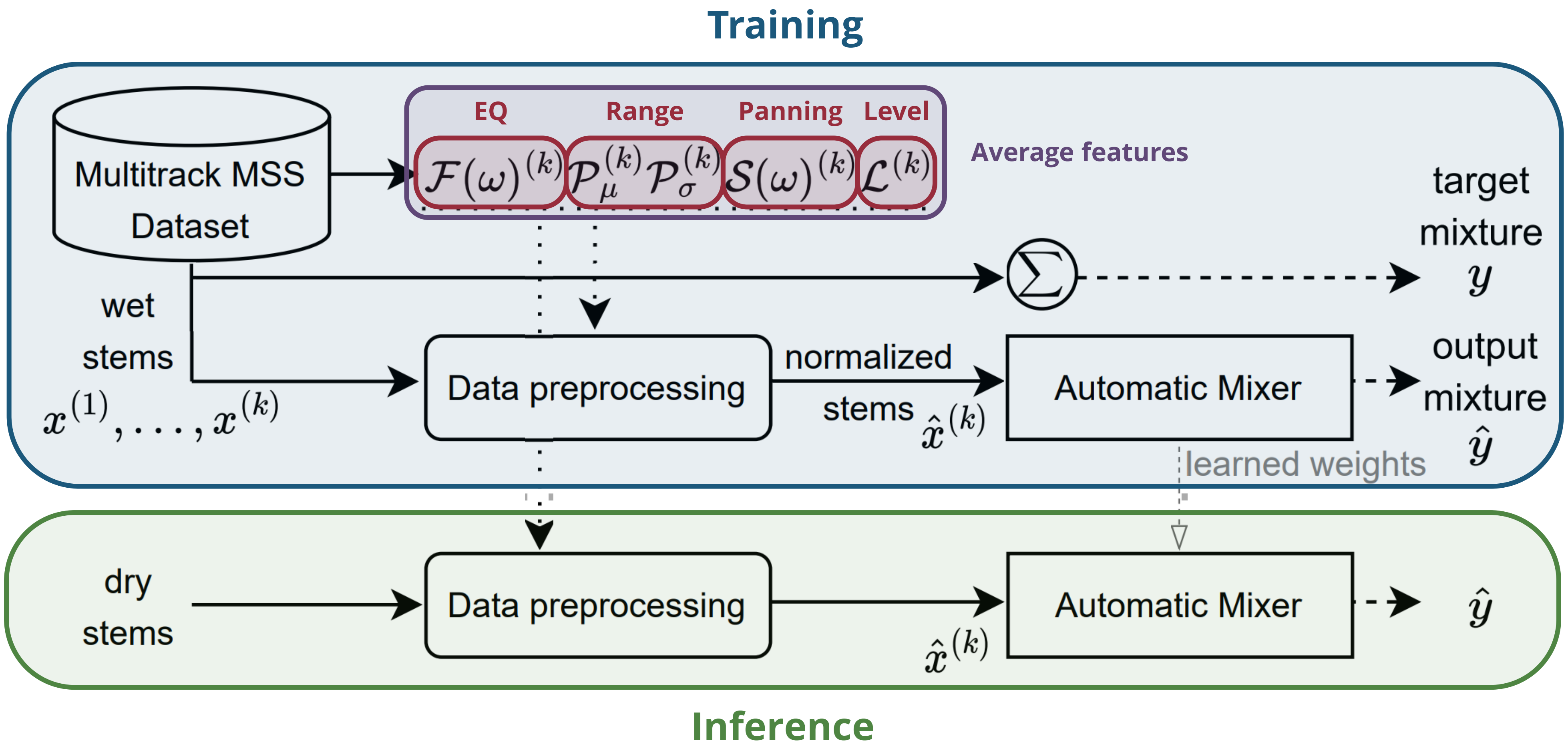


(Source: Steinmetz et al., 2022)

csteinmetz1.github.io/DeepAFx-ST

Christian J. Steinmetz, Nicholas J. Bryan, and Joshua D. Reiss, "Style Transfer of Audio Effects with Differentiable Signal Processing," *JAES*, 2022.

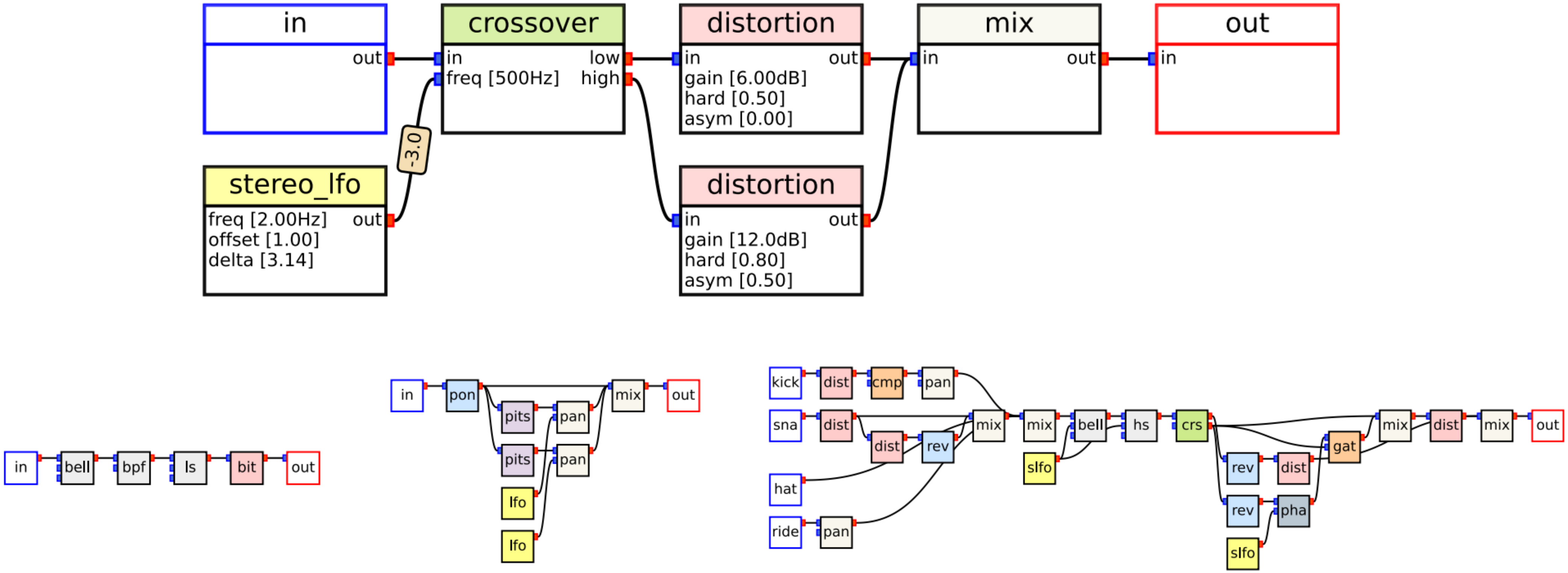
(Recap) Example: FX Normalization (Martínez-Ramírez et al., 2022)



(Source: Martínez-Ramírez et al., 2022)

Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Giorgio Fabbro, Stefan Uhlich, Chihiro Nagashima, and Yuki Mitsufuji, "Automatic music mixing with deep learning and out-of-domain data," *ISMIR*, 2022.

(Recap) Example: Audio Processing Graph (Lee et al., 2022)



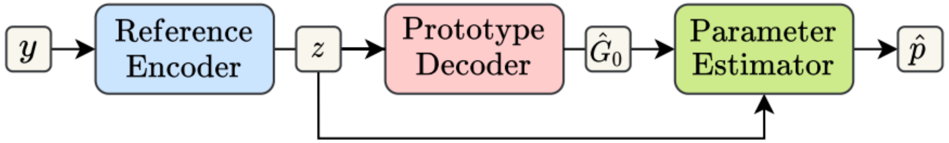
Can we predict the audio processing graph used in a reference recording?

(Source: Lee et al., 2023)

Sungho Lee, Jaehyun Park, Seungryeol Paik, and Kyogu Lee, “Blind Estimation of Audio Processing Graph,” *ICASSP*, 2023.

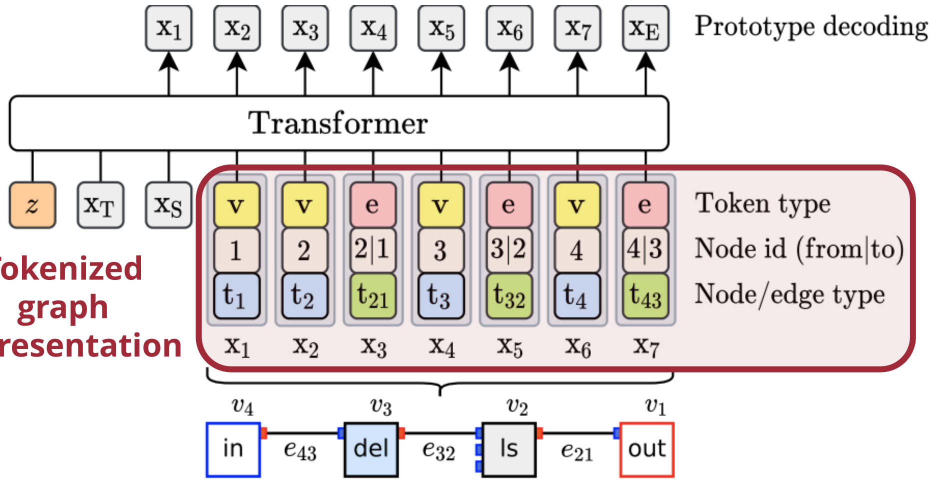
(Recap) Example: Audio Processing Graph (Lee et al., 2022)

Blind estimation framework



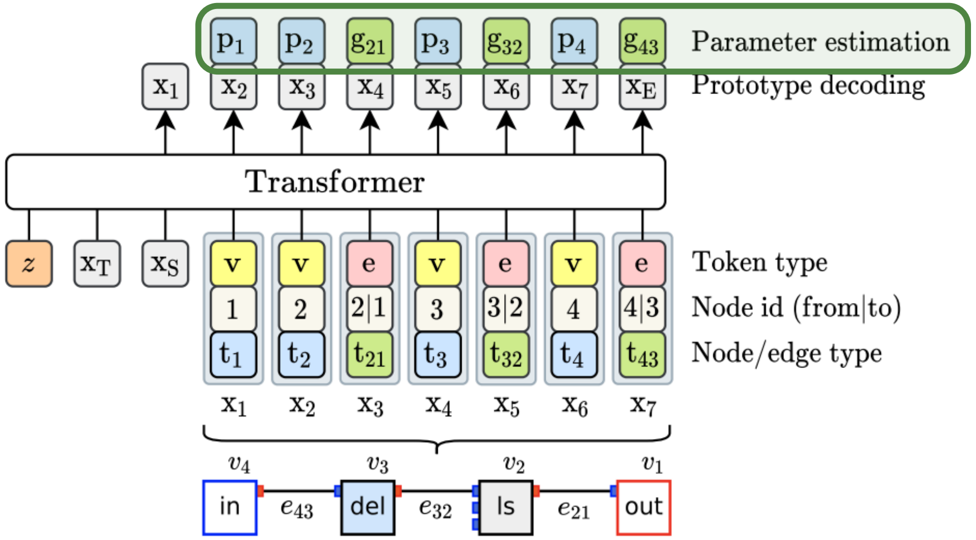
Sources are not given!

Prototype decoder


x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_E
Prototype decoding
Transformer
z x_T x_S
v v e v e v e
Token type
1 2 2|1 3 3|2 4 4|3
Node id (from|to)
t_1 t_2 t_{21} t_3 t_{32} t_4 t_{43}
Node/edge type
x_1 x_2 x_3 x_4 x_5 x_6 x_7
v_4 v_3 v_2 v_1
in e_{43} del e_{32} ls e_{21} out

Tokenized graph representation

Parameter estimator


p_1 p_2 g_{21} p_3 g_{32} p_4 g_{43}
Parameter estimation
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_E
Prototype decoding
Transformer
z x_T x_S
v v e v e v e
Token type
1 2 2|1 3 3|2 4 4|3
Node id (from|to)
t_1 t_2 t_{21} t_3 t_{32} t_4 t_{43}
Node/edge type
x_1 x_2 x_3 x_4 x_5 x_6 x_7
v_4 v_3 v_2 v_1
in e_{43} del e_{32} ls e_{21} out

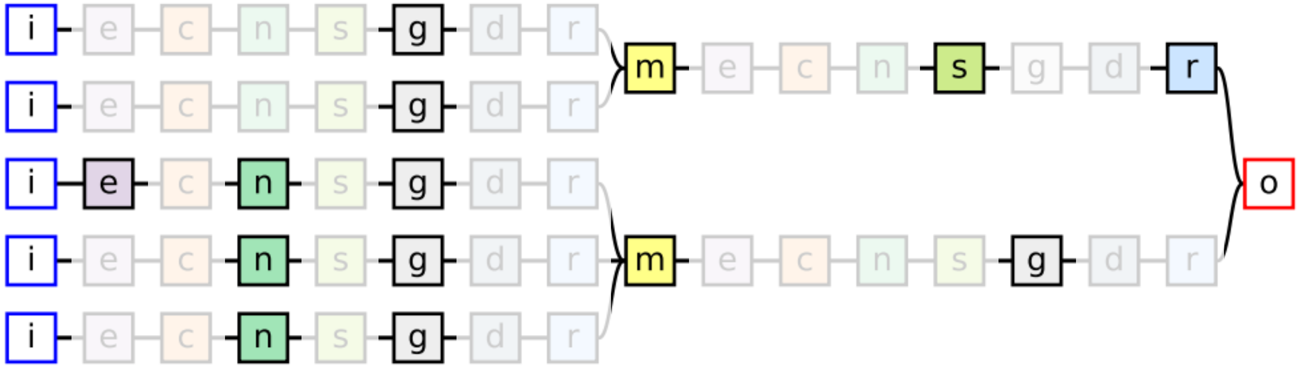
(Source: Lee et al., 2023)

Sungho Lee, Jaehyun Park, Seungryeol Paik, and Kyogu Lee, "Blind Estimation of Audio Processing Graph," *ICASSP*, 2023.

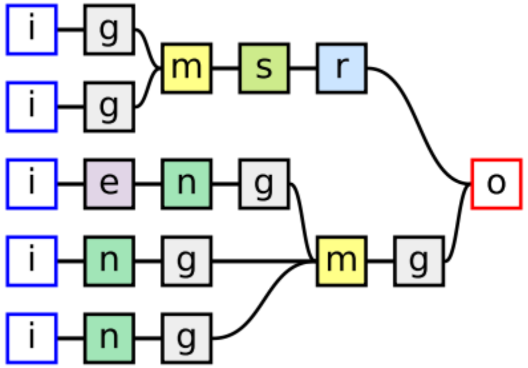
(Recap) Example: Music Mixing Graph (Lee et al., 2024)

Can we predict the music mixing graph given the sources and reference mixture?

Full mixing console (before pruning)



Pruned graph



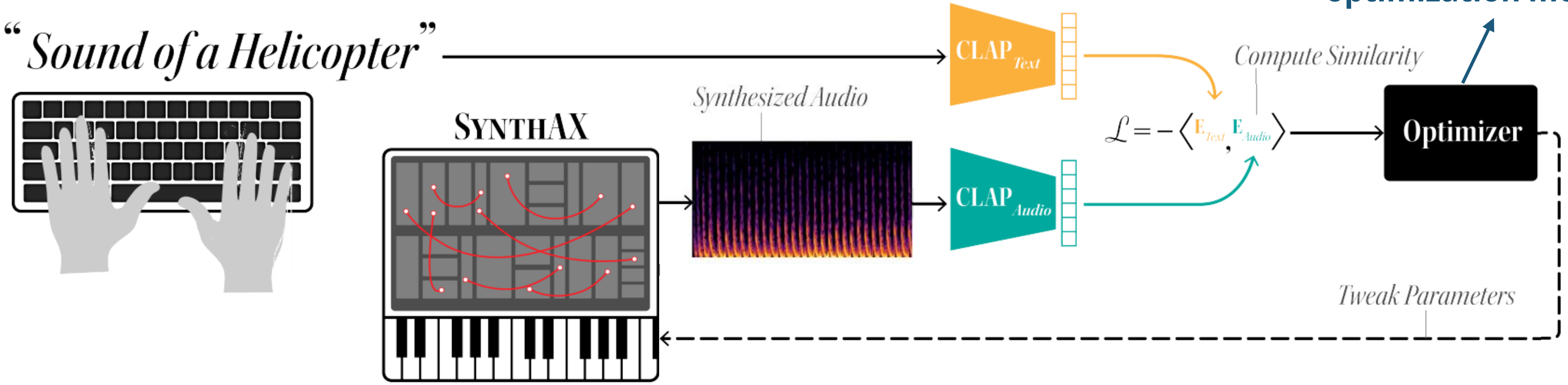
(Source: Lee et al., 2024)

sh-lee97.github.io/grafx-prune

Sungho Lee, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Stefan Uhlich, Giorgio Fabbro, Kyogu Lee, and Yuki Mitsufuji, "Searching For Music Mixing Graphs: A Pruning Approach," *DAFx*, 2024.

(Recap) Example: CTAG (Cherep et al., 2024)

Non-gradient-based optimization methods



Does not need to be differentiable!

(Source: Cherep et al., 2024)

ctag.media.mit.edu

Manuel Cherep, Nikhil Singh, and Jessica Shand, "Creative Text-to-Audio Generation via Synthesizer Programming," *ICML*, 2024.

Interactive Systems

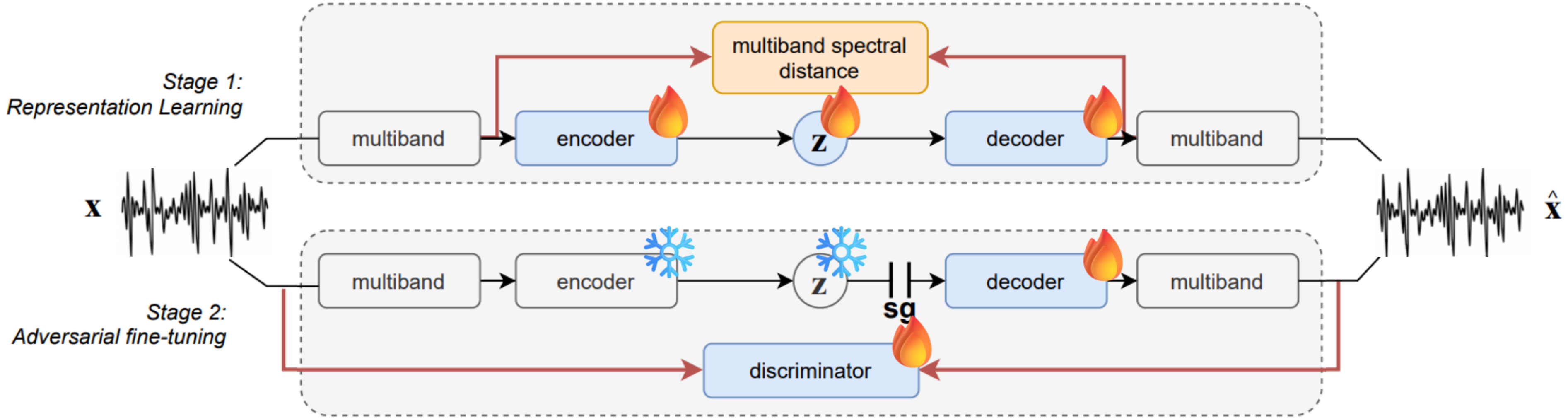
Example: RAVE (Caillon & Esling, 2022)

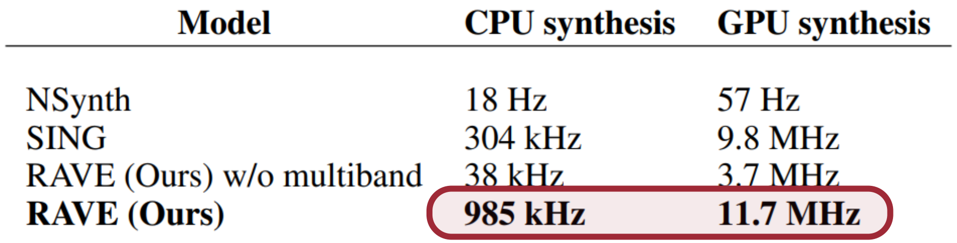
youtu.be/dMZs04TzxUI & github.com/acids-ircam/RAVE

Antoine Caillon and Philippe Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

Example: RAVE (Caillon & Esling, 2022)

16-band decomposition, 48kHz





Model	CPU synthesis	GPU synthesis
NSynth	18 Hz	57 Hz
SING	304 kHz	9.8 MHz
RAVE (Ours) w/o multiband	38 kHz	3.7 MHz
RAVE (Ours)	**985 kHz**	**11.7 MHz**

Realtime capable on CPUs & GPUs

anonymous84654.github.io/RAVE_anonymous

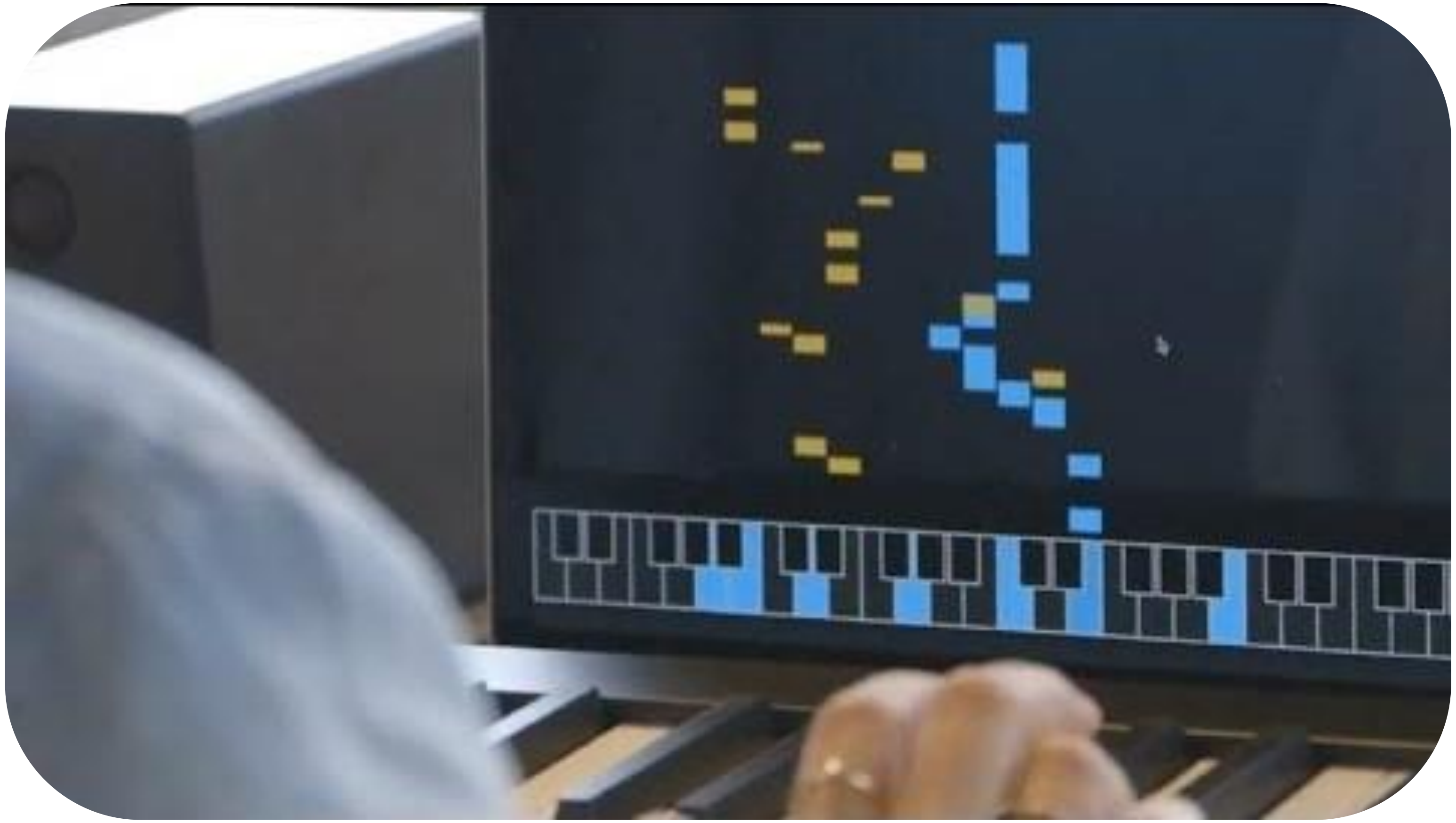
Antoine Caillon and Philippe Esling, “RAVE: A variational autoencoder for fast and high-quality neural audio synthesis,” *arXiv preprint arXiv:2111.05011*, 2021.

Example: RAVE (Caillon & Esling, 2022)

youtu.be/jAIRf4nGgYI & github.com/acids-ircam/RAVE

Antoine Caillon and Philippe Esling, “RAVE: A variational autoencoder for fast and high-quality neural audio synthesis,” *arXiv preprint arXiv:2111.05011*, 2021.

Example: A.I. Duet (Mann et al, 2016)



youtu.be/0ZE1bfPtvZo
experiments.withgoogle.com/ai/ai-duet/view

github.com/googlecreativelab/aiexperiments-ai-duet

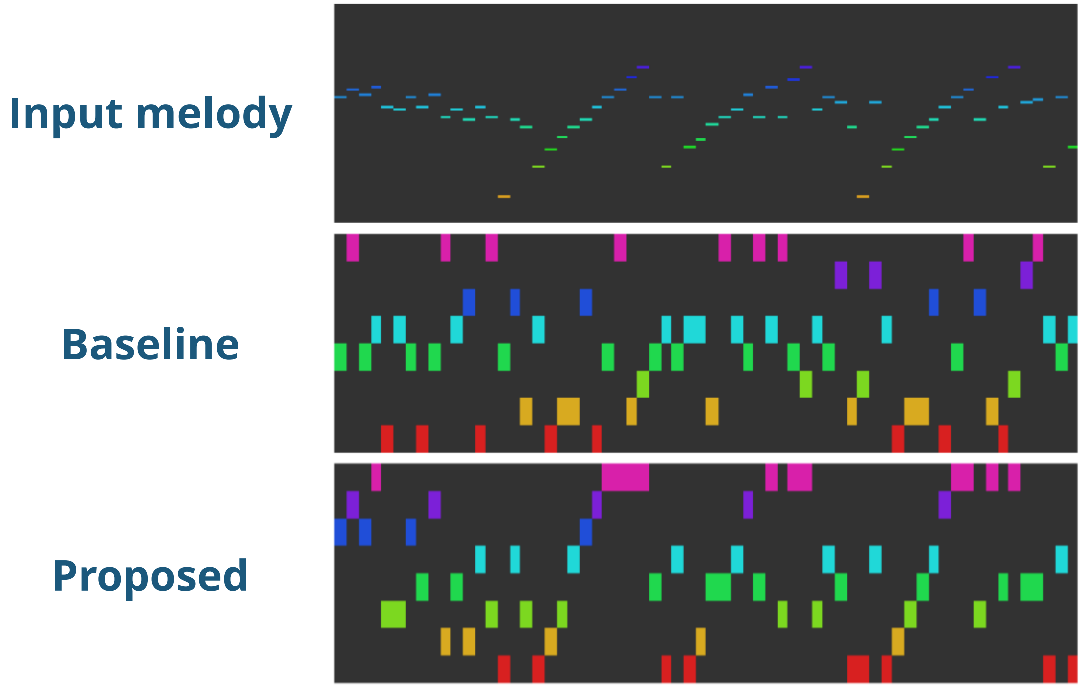
Example: Piano Genie (Donahue et al., 2018)

youtu.be/YRb0XAnUpIk & magenta.tensorflow.org/pianogenie

piano-genie.glitch.me/

Chris Donahue, Ian Simon, and Sander Dieleman, "Piano Genie," IUI, 2019.

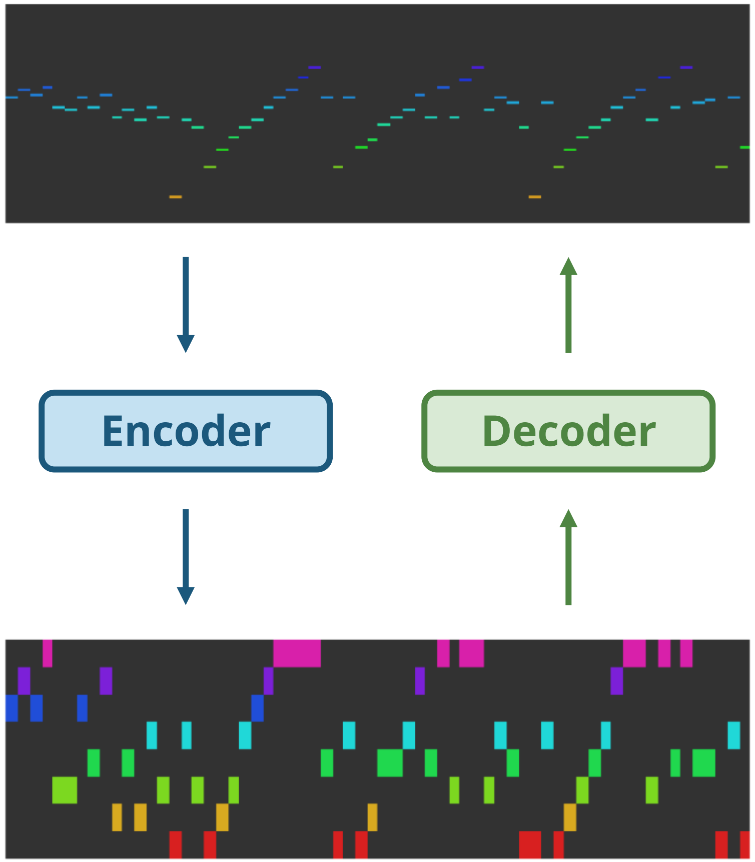
Example: Piano Genie (Donahue et al., 2018)



Input melody

Baseline

Proposed



Encoder

Decoder

(Source: Donahue et al., 2019)

Chris Donahue, Ian Simon, and Sander Dieleman, “Piano Genie,” *IUI*, 2019.

Example: Fruit Genie (2019)

youtu.be/HoVs4kC68no

magenta.tensorflow.org/fruitgenie

Example: Fruit Genie Live (2019)

youtu.be/L4wvXrPmIkU

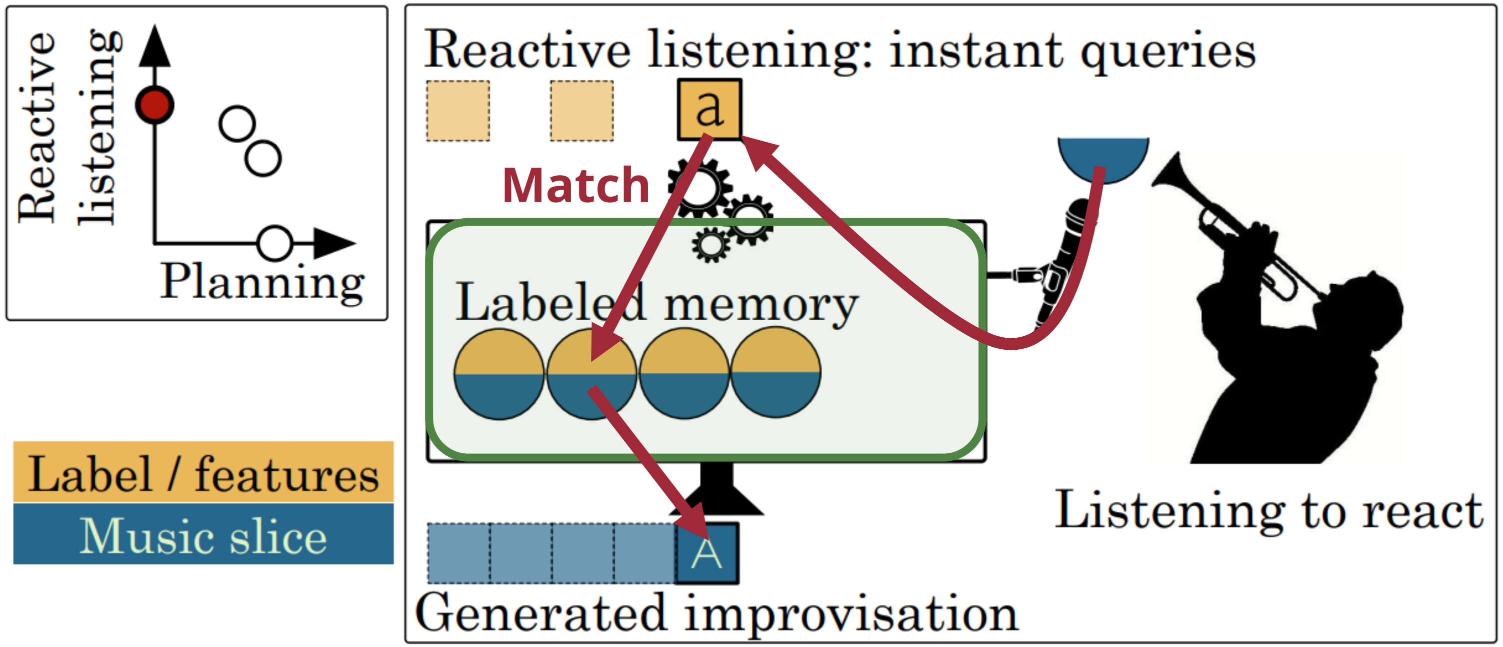
magenta.tensorflow.org/fruitgenie

Example: AI Creative Agents (2015)

On the imposed theme of "The Man I Love", which Piaf and Schwarzkopf never sang, the creative agents "improvise" from the voices of these stars, adapting to the harmony and tempo in real-time.

youtu.be/DggF9m9xqik & github.com/DYCI2/Dicy2

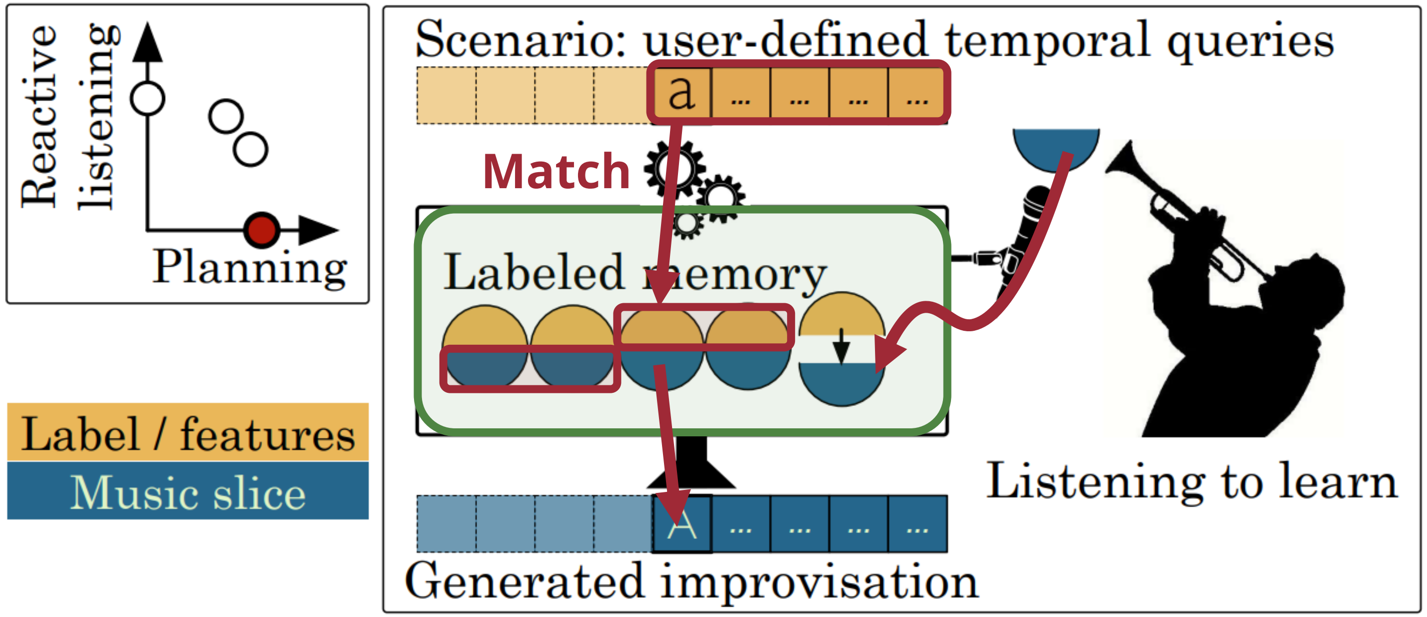
Example: Somax 2 (Nika et al., 2017)



(Source: Nika et al., 2017)

Jérôme Nika, Ken Déguernel, Axel Chemla–Romeu-Santos, Emmanuel Vincent, and Gérard Assayag, “DYCI2 agents: merging the ”free”, ”reactive”, and ”scenario-based” music generation paradigms,” *ICMC*, 2017.

Example: ImproteK (Nika et al., 2017)



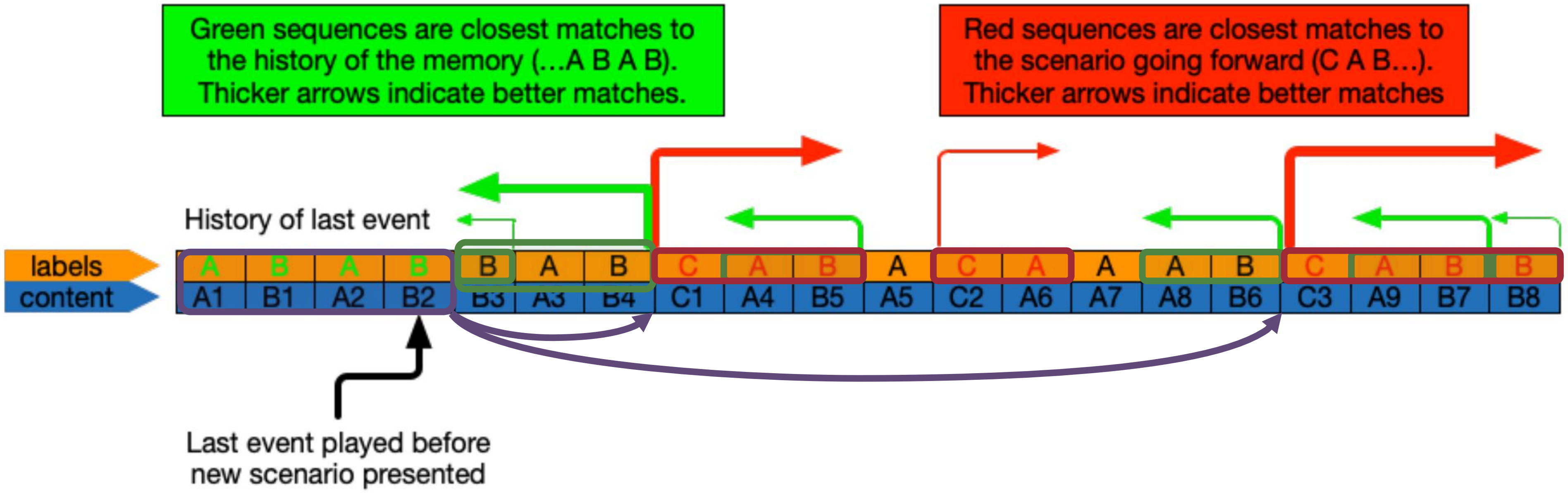
(Source: Nika et al., 2017)

Jérôme Nika, Ken Déguernel, Axel Chemla–Romeu-Santos, Emmanuel Vincent, and Gérard Assayag, "DYCI2 agents: merging the "free", "reactive", and "scenario-based" music generation paradigms," *ICMC*, 2017.

Example: ImproteK (Nika et al., 2017)

For the scenario C A B B C C B A:

Matching both the history of the memory and the future of the scenario



Green sequences are closest matches to the history of the memory (...A B A B). Thicker arrows indicate better matches.

Red sequences are closest matches to the scenario going forward (C A B...). Thicker arrows indicate better matches

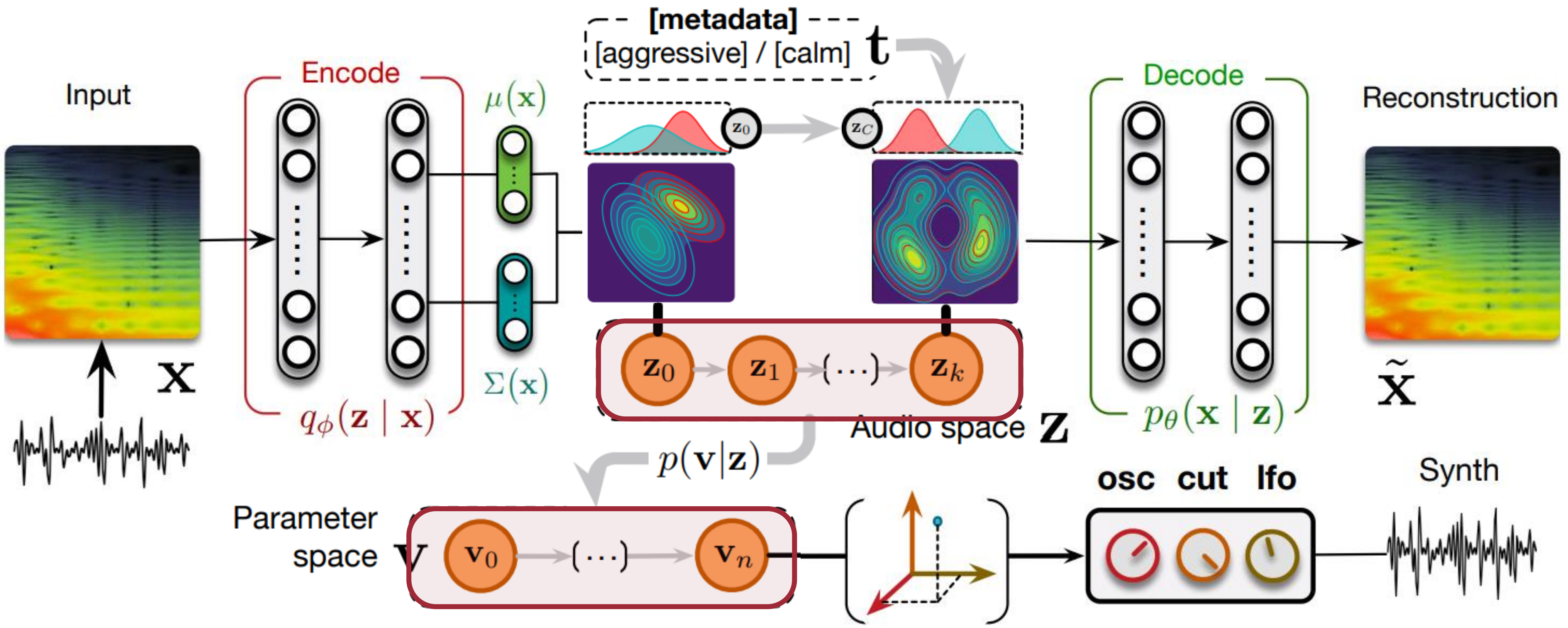
labels	A	B	A	B	B	A	B	C	A	B	A	C	A	A	A	B	C	A	B	B
content	A1	B1	A2	B2	B3	A3	B4	C1	A4	B5	A5	C2	A6	A7	A8	B6	C3	A9	B7	B8

Last event played before new scenario presented

(Source: Nika et al., 2017)

Jérôme Nika, Augustin Muller, Joakim Borg, Gérard Assayag, and Matthew Ostrowski, “Dicy2 for Live,” *Ircam*, 2022.

Example: FlowSynth (Esling et al., 2019)



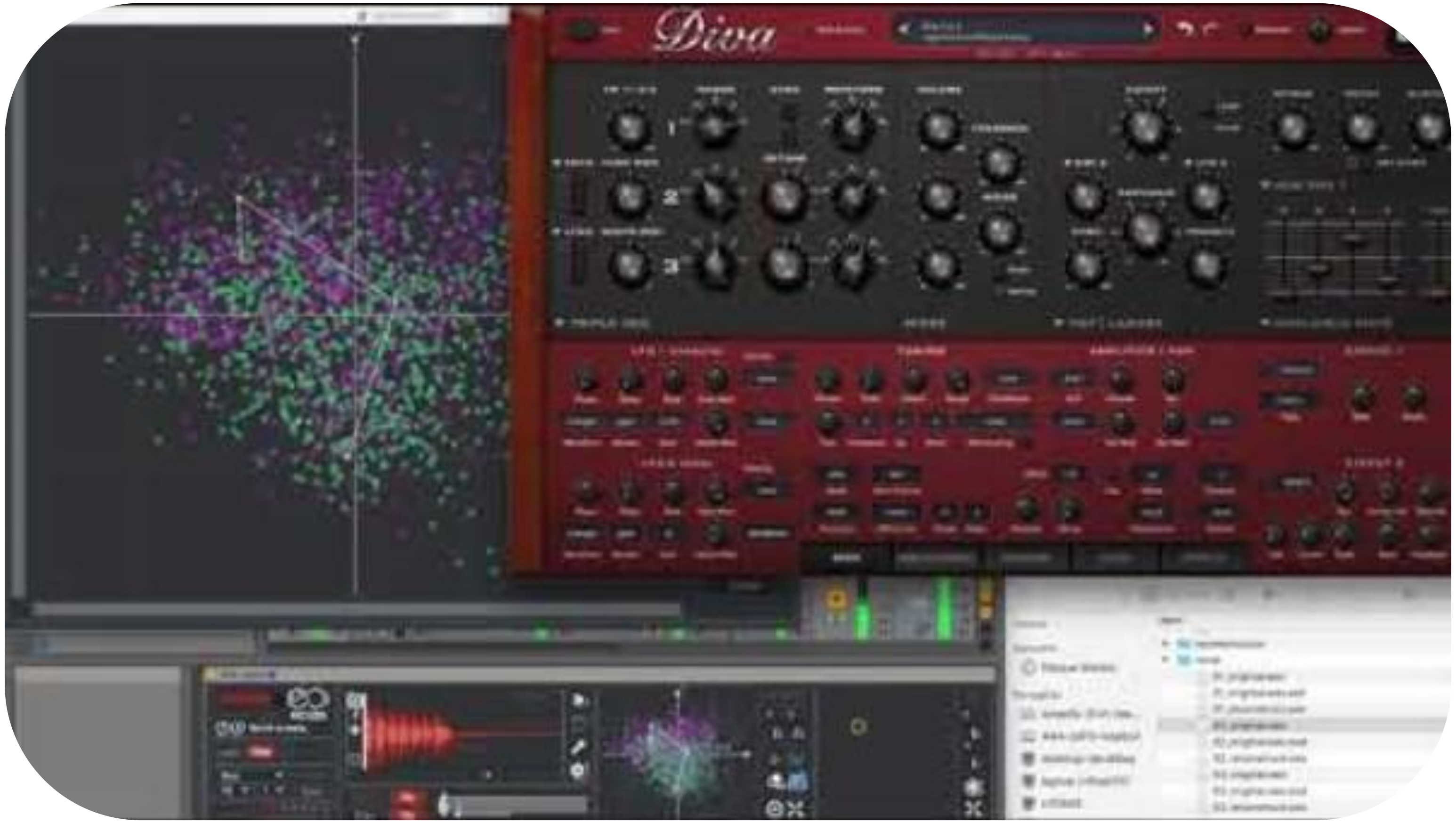
Normalizing flows models

(another class of generative models)

(Source: Esling et al., 2019)

Philippe Esling, Naotake Masuda, Adrien Bardet, Romeo Despres, and Axel Chemla–Romeu-Santos, "Universal audio synthesizer control with normalizing flows," *DaFX*, 2019.

Example: FlowSynth (Esling et al., 2019)

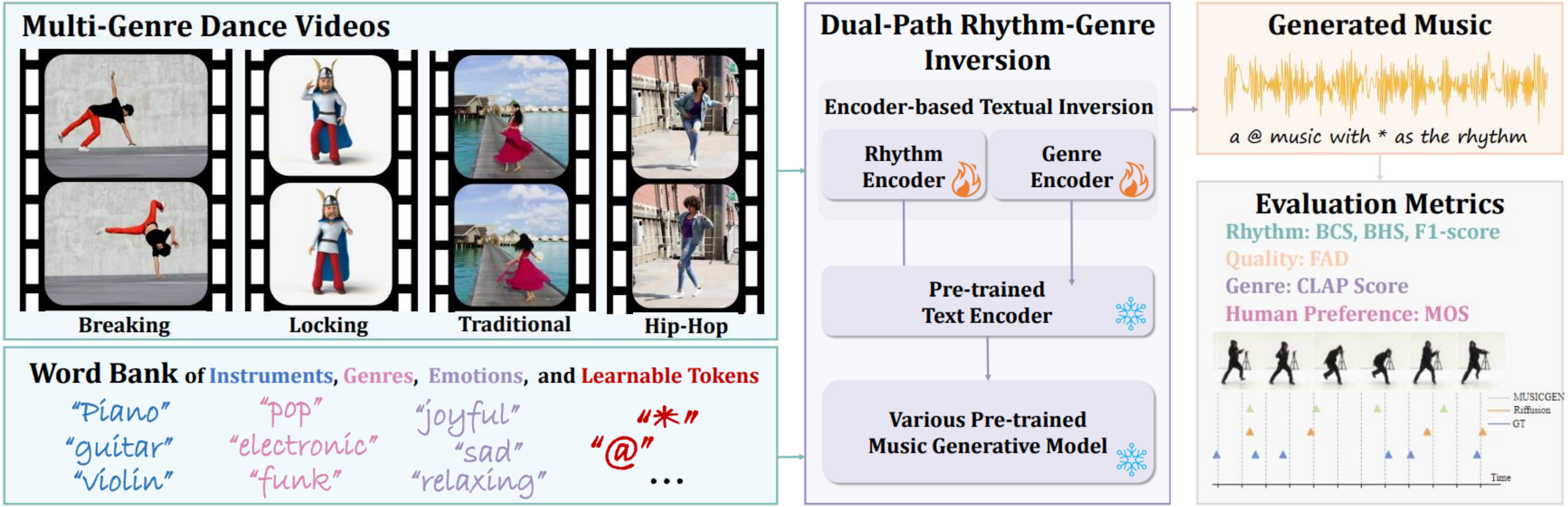


youtu.be/UufQwUitBIw

Philippe Esling, Naotake Masuda, Adrien Bardet, Romeo Despres, and Axel Chemla–Romeu-Santos, “Universal audio synthesizer control with normalizing flows,” *DaFX*, 2019.

Multimodal Systems

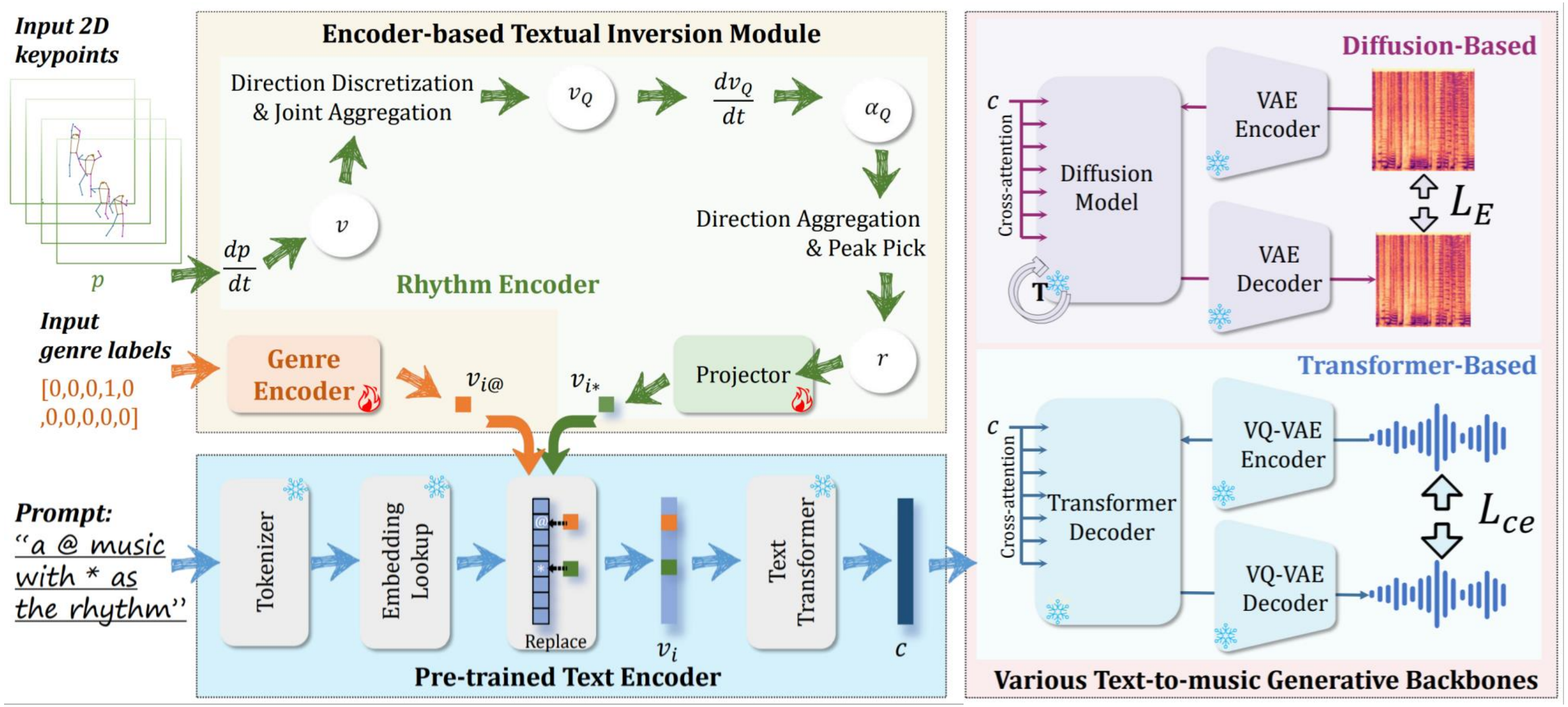
Example: Dance-to-music Generation (Li et al., 2024)



(Source: Li et al., 2024)

Sifei Li, Weiming Dong, Yuxin Zhang, Fan Tang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu, "Dance-to-Music Generation with Encoder-based Textual Inversion," *SIGGRAPH ASIA*, 2024.

Example: Dance-to-music Generation (Li et al., 2024)



(Source: Li et al., 2024)

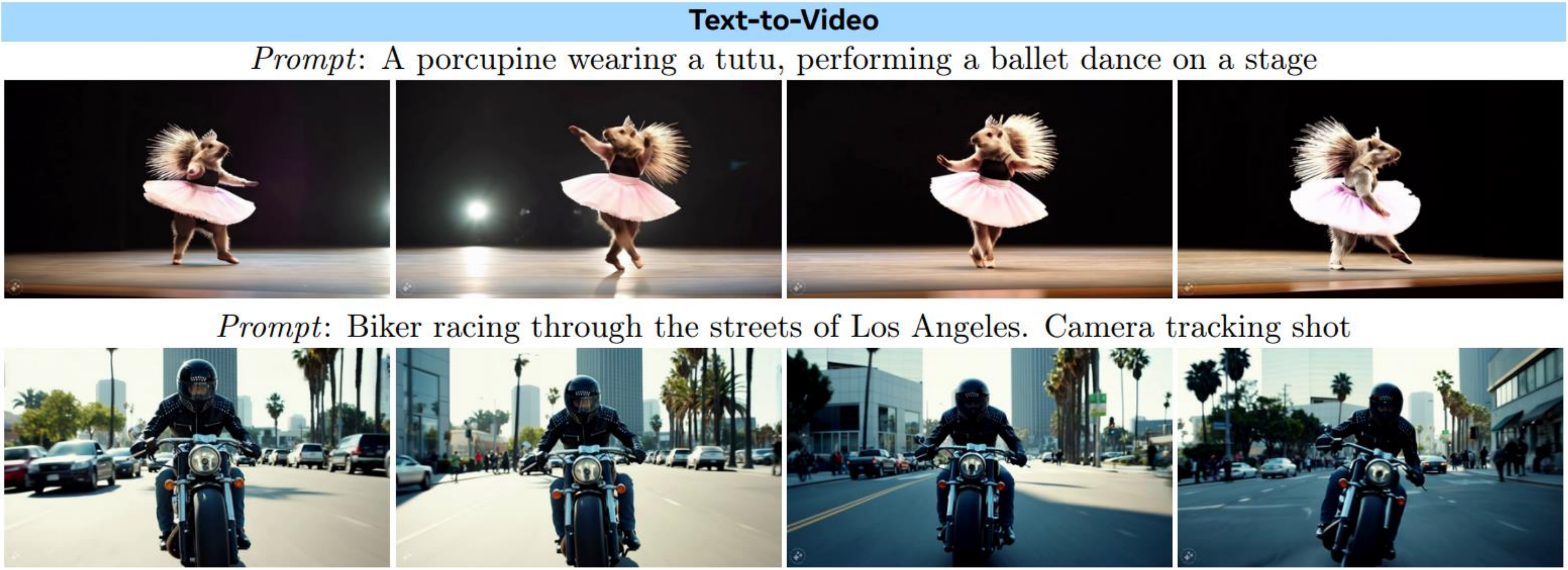
Sifei Li, Weiming Dong, Yuxin Zhang, Fan Tang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu, "Dance-to-Music Generation with Encoder-based Textual Inversion," *SIGGRAPH ASIA*, 2024.

Example: Dance-to-music Generation (Li et al., 2024)

youtu.be/y2pG2S5xDLY

Sifei Li, Weiming Dong, Yuxin Zhang, Fan Tang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu, "Dance-to-Music Generation with Encoder-based Textual Inversion," *SIGGRAPH ASIA*, 2024.

Example: MovieGen (2024)



Text-to-Video

Prompt: A porcupine wearing a tutu, performing a ballet dance on a stage

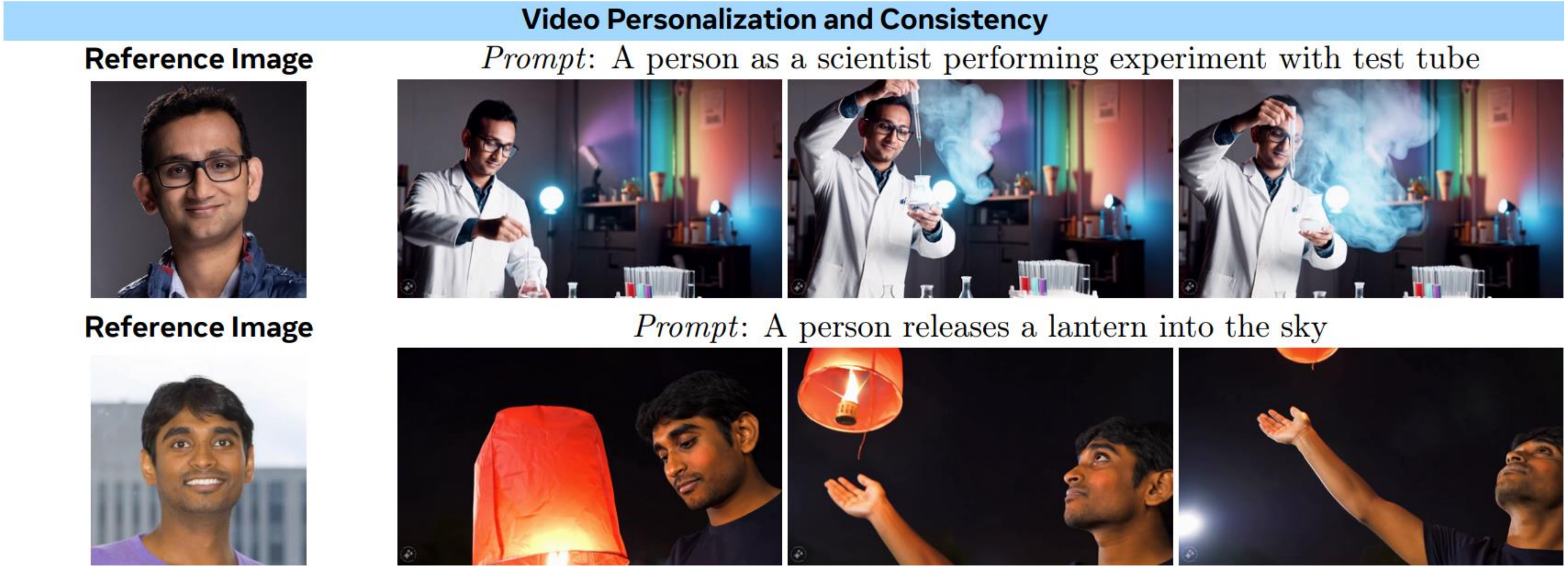
Prompt: Biker racing through the streets of Los Angeles. Camera tracking shot

(Source: Movie Gen Team, 2024)

ai.meta.com/research/movie-gen/

Movie Gen Team, "Movie Gen: A Cast of Media Foundation Models," *arXiv preprint arXiv:2410.13720*, 2024.

Example: MovieGen (2024)



Prompt: A person as a scientist performing experiment with test tube

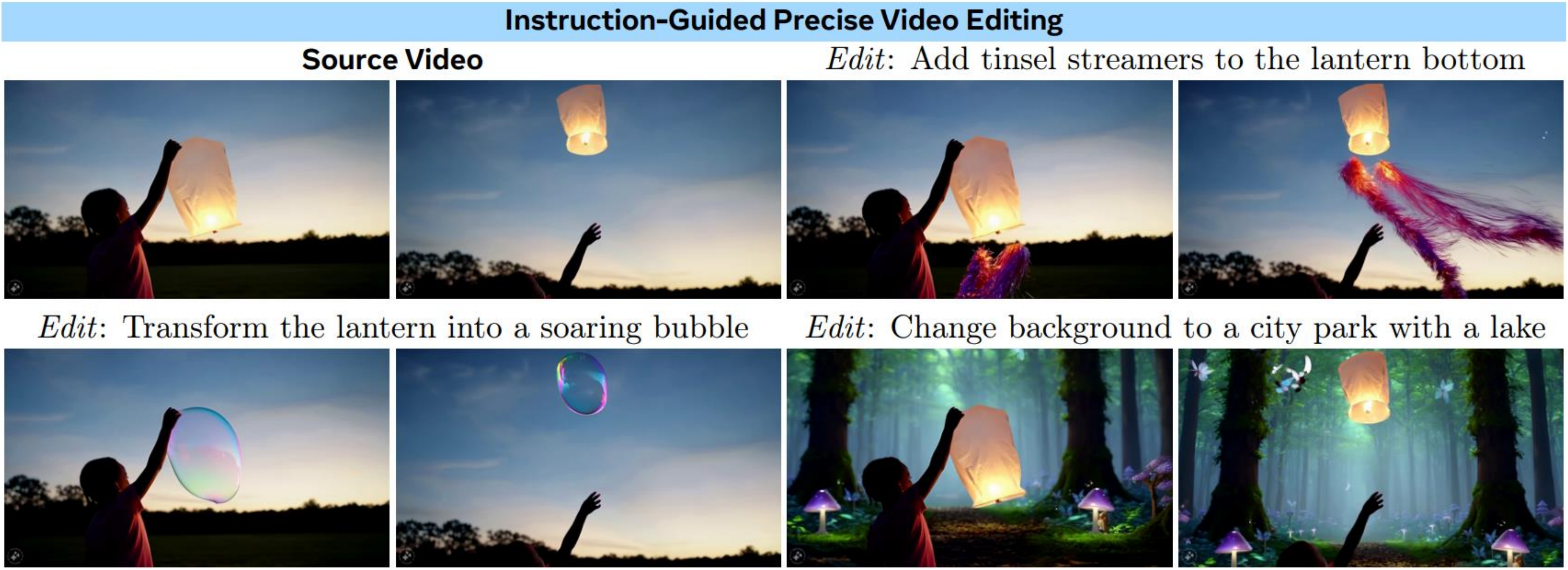
Prompt: A person releases a lantern into the sky

(Source: Movie Gen Team, 2024)

ai.meta.com/research/movie-gen/

Movie Gen Team, "Movie Gen: A Cast of Media Foundation Models," *arXiv preprint arXiv:2410.13720*, 2024.

Example: MovieGen (2024)

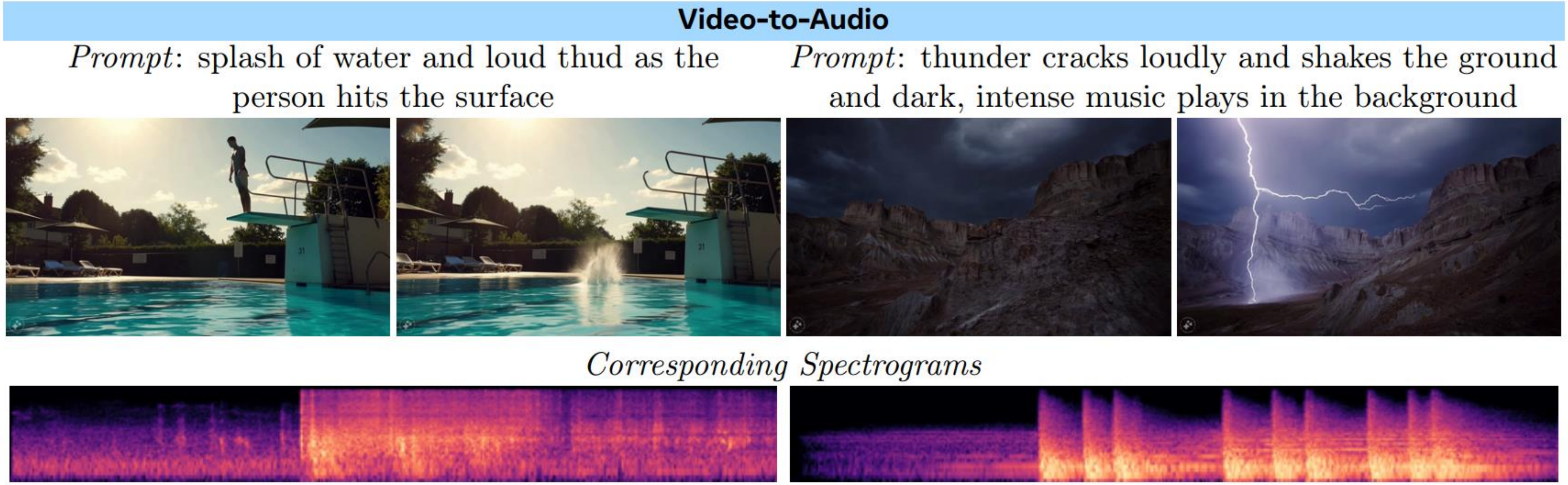

Instruction-Guided Precise Video Editing
Source Video
Edit: Add tinsel streamers to the lantern bottom
Edit: Transform the lantern into a soaring bubble
Edit: Change background to a city park with a lake

(Source: Movie Gen Team, 2024)

ai.meta.com/research/movie-gen/

Movie Gen Team, "Movie Gen: A Cast of Media Foundation Models," *arXiv preprint arXiv:2410.13720*, 2024.

Example: MovieGen (2024)

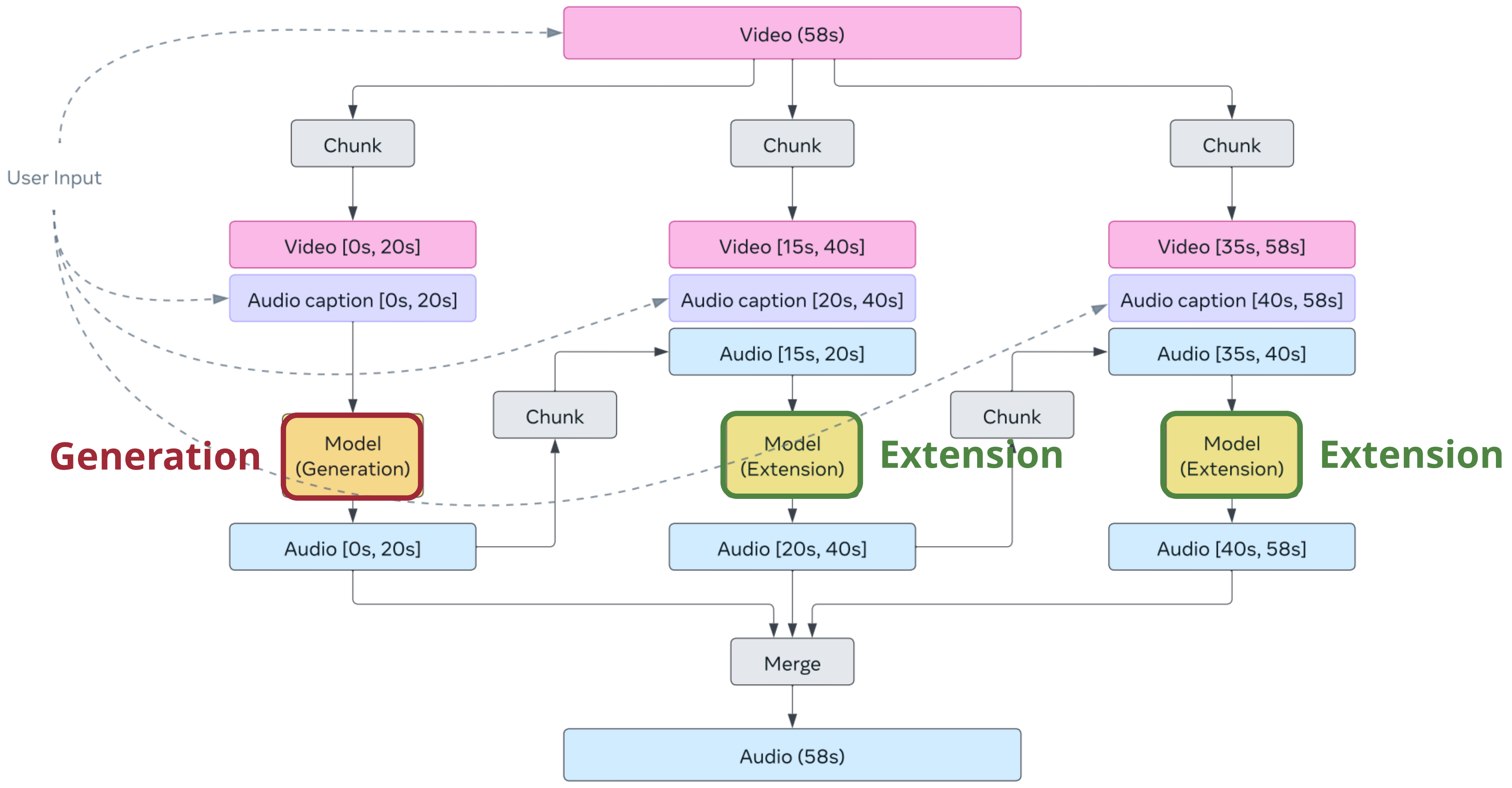

Video-to-Audio
Prompt: splash of water and loud thud as the person hits the surface
Prompt: thunder cracks loudly and shakes the ground and dark, intense music plays in the background
Corresponding Spectrograms

(Source: Movie Gen Team, 2024)

ai.meta.com/research/movie-gen/

Movie Gen Team, “Movie Gen: A Cast of Media Foundation Models,” *arXiv preprint arXiv:2410.13720*, 2024.

Example: MovieGen (2024)

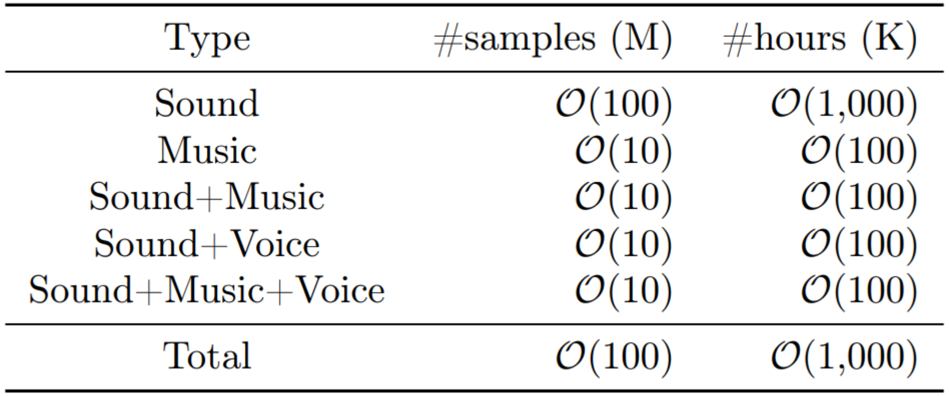


(Source: Movie Gen Team, 2024)

Movie Gen Team, "Movie Gen: A Cast of Media Foundation Models," *arXiv preprint arXiv:2410.13720*, 2024.

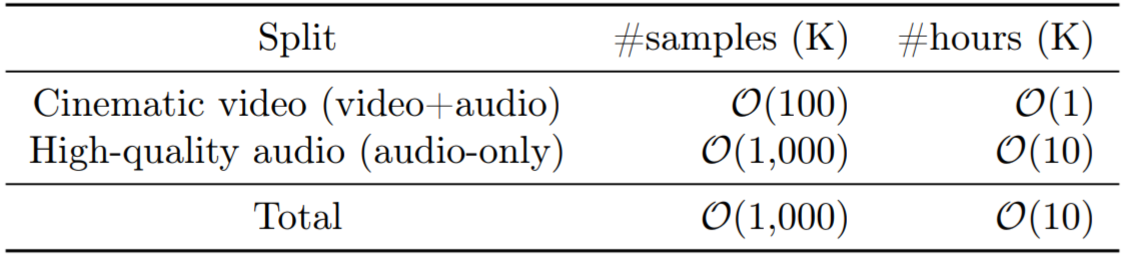
Example: MovieGen (2024)

Pretraining Data



Type	#samples (M)	#hours (K)
Sound	$\mathcal{O}(100)$	$\mathcal{O}(1{,}000)$
Music	$\mathcal{O}(10)$	$\mathcal{O}(100)$
Sound+Music	$\mathcal{O}(10)$	$\mathcal{O}(100)$
Sound+Voice	$\mathcal{O}(10)$	$\mathcal{O}(100)$
Sound+Music+Voice	$\mathcal{O}(10)$	$\mathcal{O}(100)$
Total	$\mathcal{O}(100)$	$\mathcal{O}(1{,}000)$

Finetuning Data



Split	#samples (K)	#hours (K)
Cinematic video (video+audio)	$\mathcal{O}(100)$	$\mathcal{O}(1)$
High-quality audio (audio-only)	$\mathcal{O}(1{,}000)$	$\mathcal{O}(10)$
Total	$\mathcal{O}(1{,}000)$	$\mathcal{O}(10)$

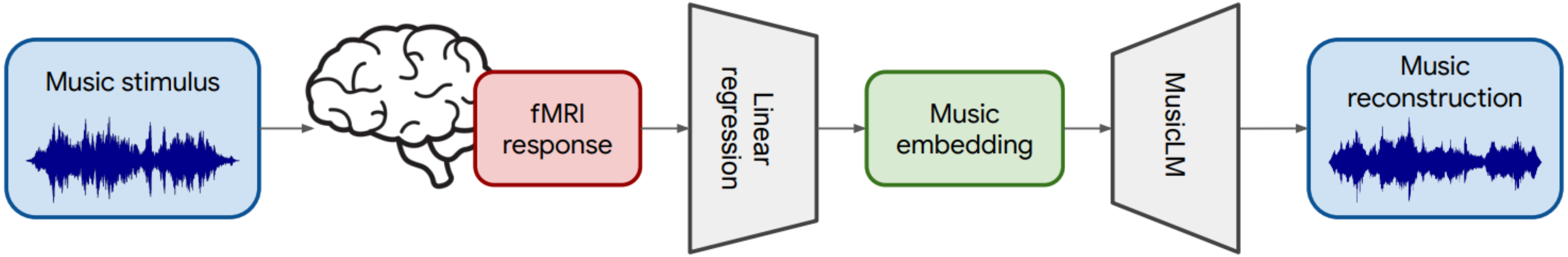
(Source: Movie Gen Team, 2024)

Pretrained on >1000K hr audio
Finetuned on >1K hr cinematic videos & >10K hr HQ audio

Movie Gen Team, “Movie Gen: A Cast of Media Foundation Models,” *arXiv preprint arXiv:2410.13720*, 2024.

Example: Brain2Music (Denk et al., 2023)

Brain as the encoder!

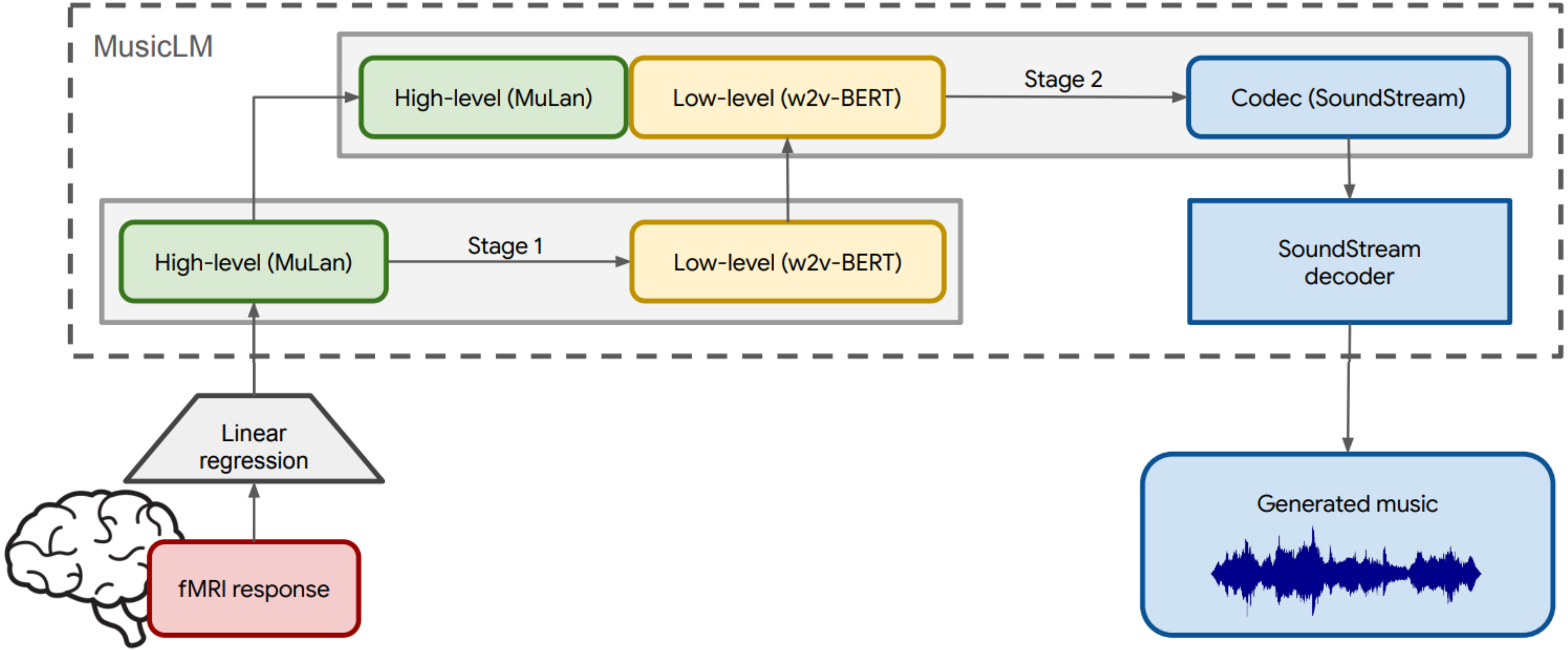


(Source: Denk et al., 2023)

Can we decode human brain-encoded music?

Timo I. Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto, “Brain2Music: Reconstructing Music from Human Brain Activity,” *arXiv preprint arXiv:2307.11078*, 2023.

Example: Brain2Music (Denk et al., 2023)



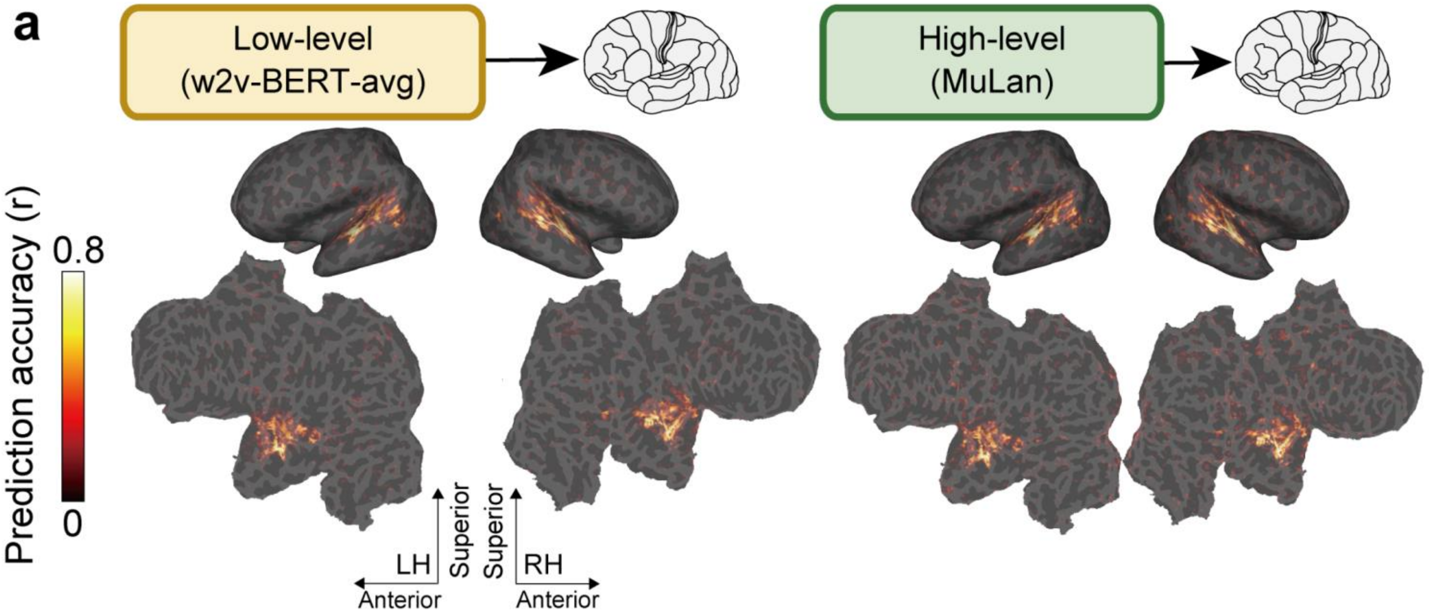
(Source: Denk et al., 2023)

google-research.github.io/seanet/brain2music

Timo I. Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto, “Brain2Music: Reconstructing Music from Human Brain Activity,” *arXiv preprint arXiv:2307.11078*, 2023.

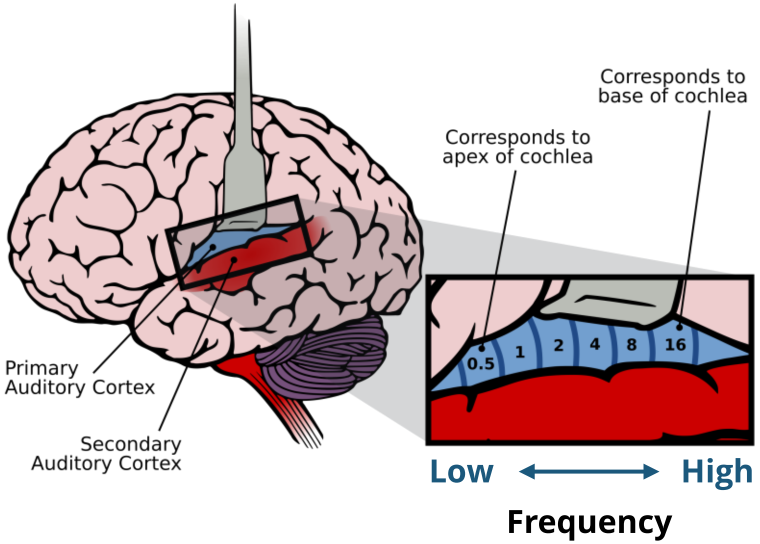
Example: Brain2Music (Denk et al., 2023)

Audio embedding to brain activity prediction



(Source: Denk et al., 2023)

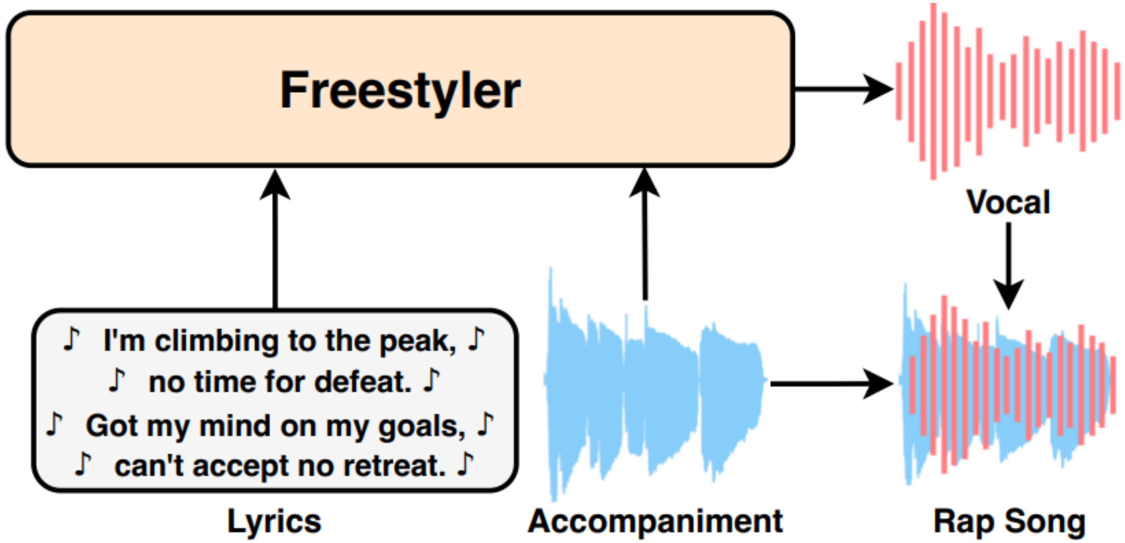
Auditory cortex



(Source: Wikimedia Commons)

Chittka L, Brockmann, CC BY-SA 2.5, via Wikimedia Commons
Timo I. Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto, "Brain2Music: Reconstructing Music from Human Brain Activity," *arXiv preprint arXiv:2307.11078*, 2023.

Example: Freestyler (Ning et al., 2024)


Freestyler
Vocal
♪ I'm climbing to the peak, ♪
♪ no time for defeat. ♪
♪ Got my mind on my goals, ♪
♪ can't accept no retreat. ♪
Lyrics
Accompaniment
Rap Song

nzqian.github.io/Freestyler

Ziqian Ning, Shuai Wang, Yuepeng Jiang, Jixun Yao, Lei He, Shifeng Pan, Jie Ding, and Lei Xie, “Drop the beat! Freestyler for Accompaniment Conditioned Rapping Voice Generation,” *arXiv preprint arXiv:2408.15474*, 2024.