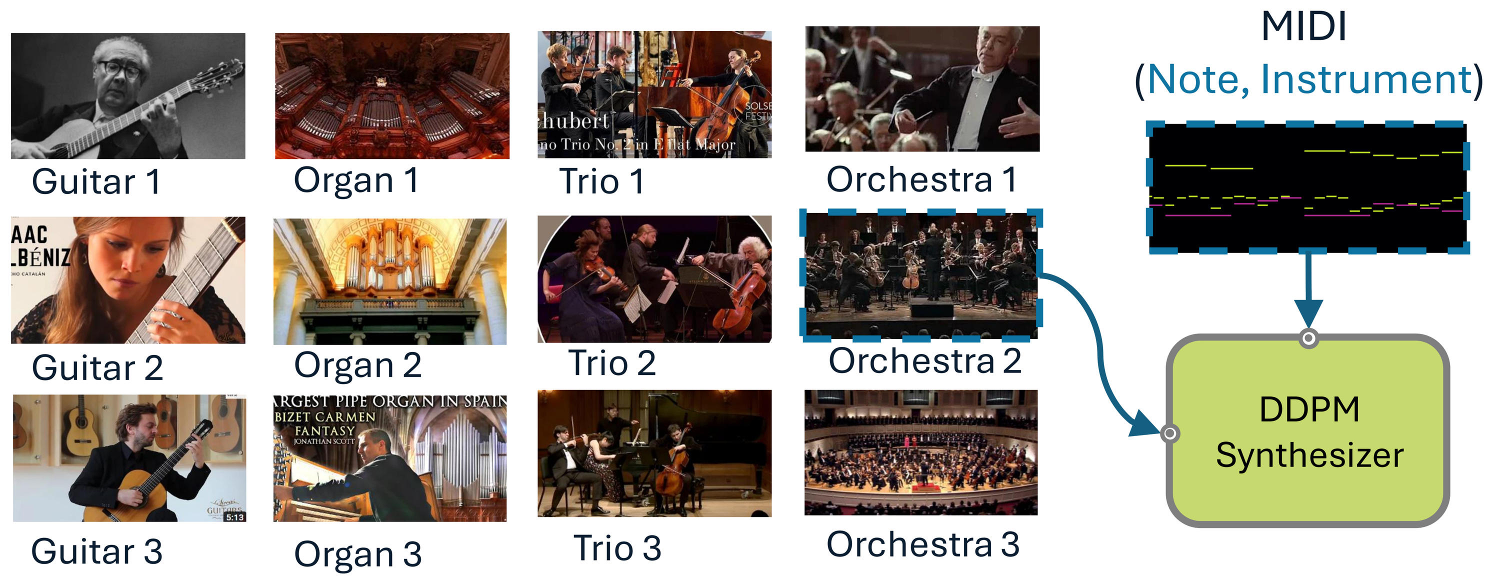
Performance Conditioning for Diffusion-Based Multi-Instrument Music Synthesis

Ben Maman (Tel Aviv University), Johannes Zeitler (International Audio Laboratories Erlangen, Germany), Meinard Müller (International Audio Laboratories Erlangen, Germany), Amit H.Bermano (Tel Aviv University)

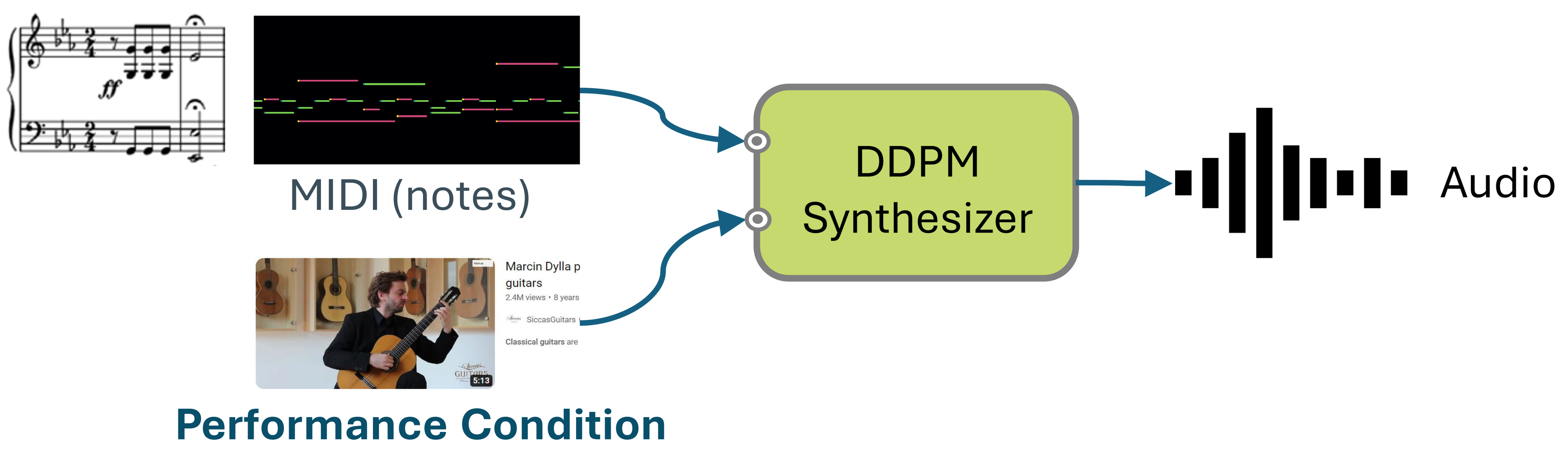
Generating multi-instrument music from symbolic music representations is an important task in Music Information Retrieval (MIR). A central but still largely unsolved problem in this context is musically and acoustically informed control in the generation process. As the main contribution of this work, we propose enhancing control of multi-instrument synthesis by conditioning a generative model on a specific performance and recording environment, thus allowing for better guidance of timbre and style. Building on state-of-the-art diffusion-based music generative models, we introduce performance conditioning – a simple tool indicating the generative model to synthesize music with style and timbre of specific instruments taken from specific performances. Our prototype is evaluated using uncurated performances with diverse instrumentation and achieves stateof-the-art FAD realism scores while allowing novel timbre and style control. Our project page, including samples and demonstrations, is available at benadar293.github.io/midipm.

Conditioning – notes, instruments, **performance ID**



Generate musical performances, control over notes, instrumentation, & acoustics

Performance Conditioning - control acoustics, timbre, recording environment…



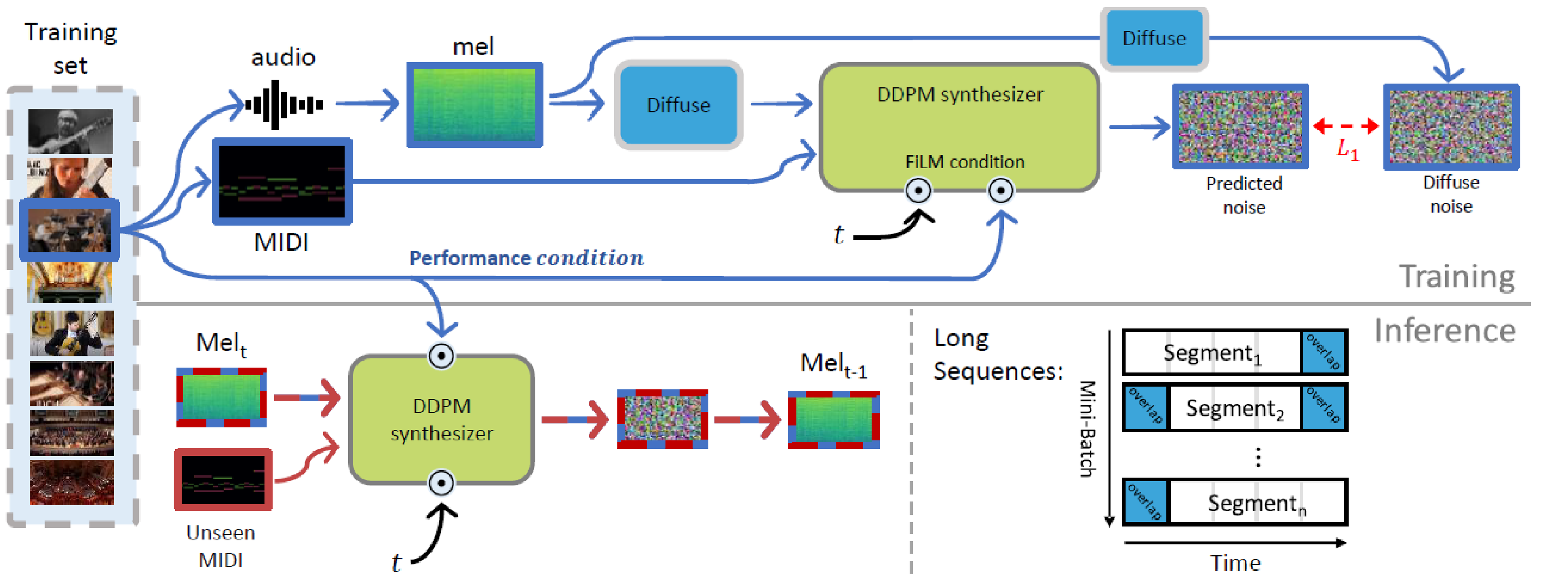
	Multi Inst.	Perf. Cond.	Orch.	Data	Real%
Maestro	✗	✓	✗	∼140H	100%
PerformanceNet	✗	✗	✗	∼1H	100%
Kim et al. 2019	✗	✗	✗	∼1H	0%
DeepPerformer	✗	✗	✗	∼1H	100%
MIDI-DDSP	✗	✗	✗	∼3H	100%
Google Magenta	✓	✗	✗	∼1500H	∼ 2%
Ours	✓	✓	✓	∼58H	100%

Improvements upon existing methods:

- Generate multi-instrument performances, including orchestra
- Train on real uncurated data using SOTA alignment
- Acoustic control using performance ID

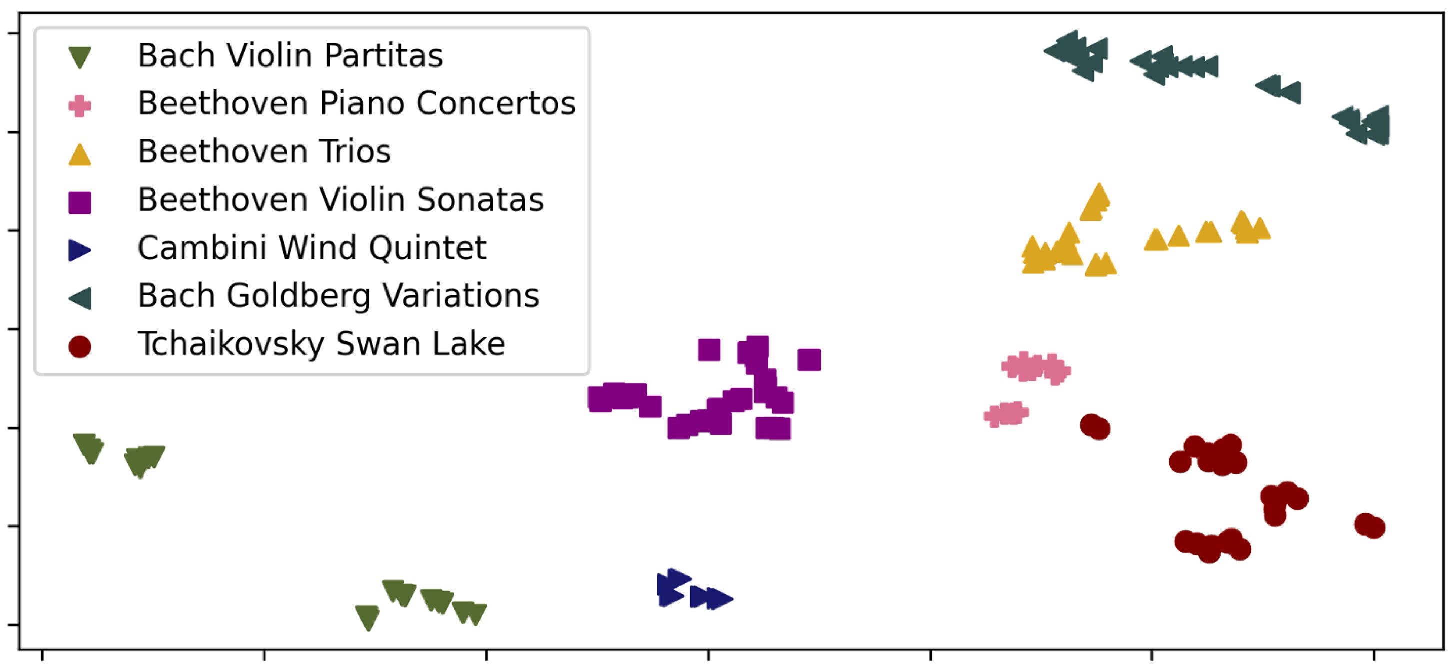
Input: Sequence of note evens (MIDI), & target performance (e.g., Karajan's 1962 recording of Beethoven's 7th Symphony)

Output: Musical performance of the new given notes in the acoustics and style of the target performance



Enahance SOTA spectrogram diffusion models:

- Use SOTA multi-instrument transcription & alignment for **real uncurated data**
  ➔ better realism than synthetic data
- Novel control (performance conditioning) – acoustics, style…
  ➔ different guitars, orchestras…
- Overlapped generation
  ➔ consistent segments, smooth transitions (inference only)



Audio tracks of same performance nicely clustered in TRILL embedding space

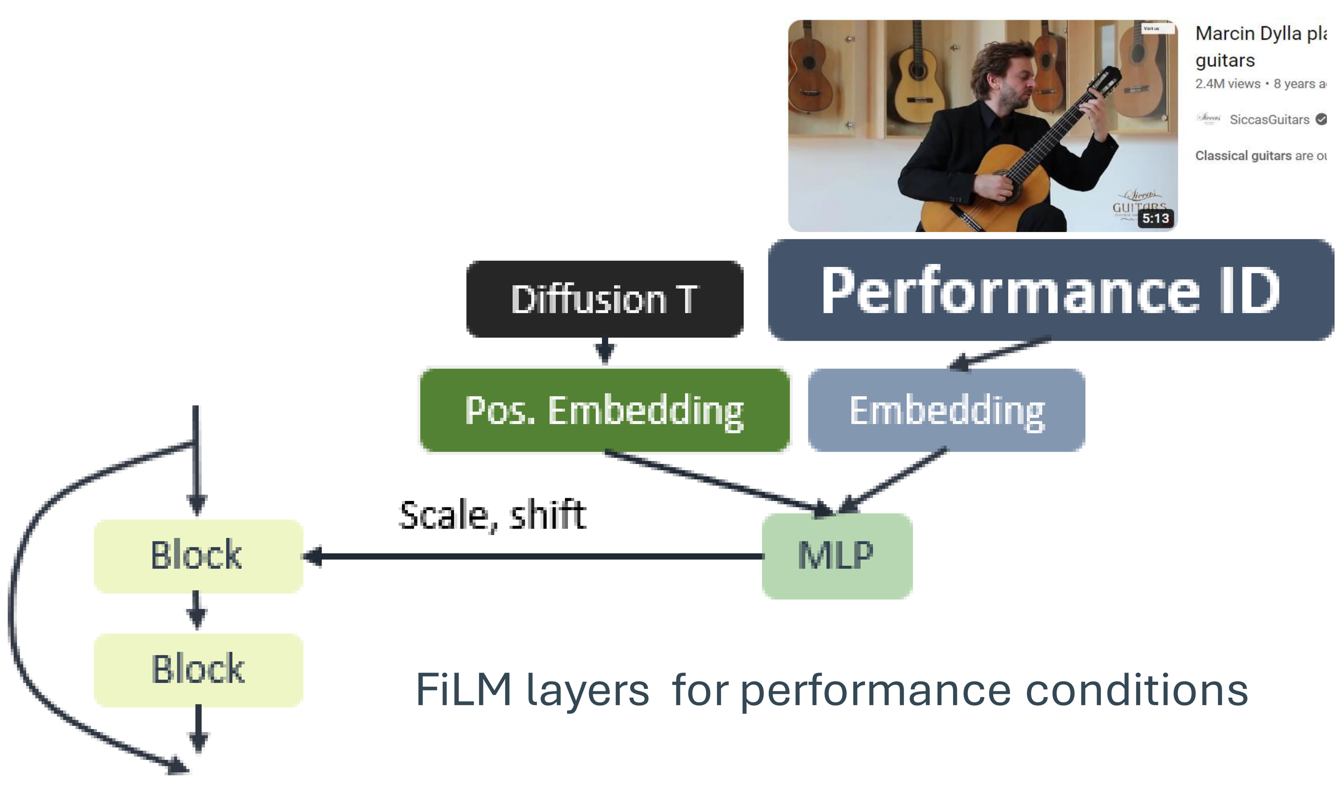
Fréchet Audio Distance (FAD):

- Group-FAD – similarity to target performance
- All-FAD – general similarity to real data

Transcription metrics

	Group-FAD↓				Perf. Acc.% Top-1/3↑	
	VGGish		TRILL			
P Con.	w/o	w/	w/o	w/	w/o	w/
T5	5.8	5.5	0.43	0.35	36/60%	**68/90%**
U-Net	7.1	**5.1**	0.5	**0.33**	14/30%	56/73%

Performance conditioning dramatically improves similarity to target (Group-FAD & classification)



FiLM layers for performance conditions

Model	All-FAD↓				Transcription N/N+I↑	
	VGGish		TRILL			
P Con.	w/o	w/	w/o	w/	w/o	w/
T5	3.9	3.5	0.12	**0.09**	62/38%	**63/47%**
U-Net	**3.4**	3.9	0.12	0.11	**63/47%**	62/46%

Performance conditioning does not hurt general quality (All-FAD & transcription))