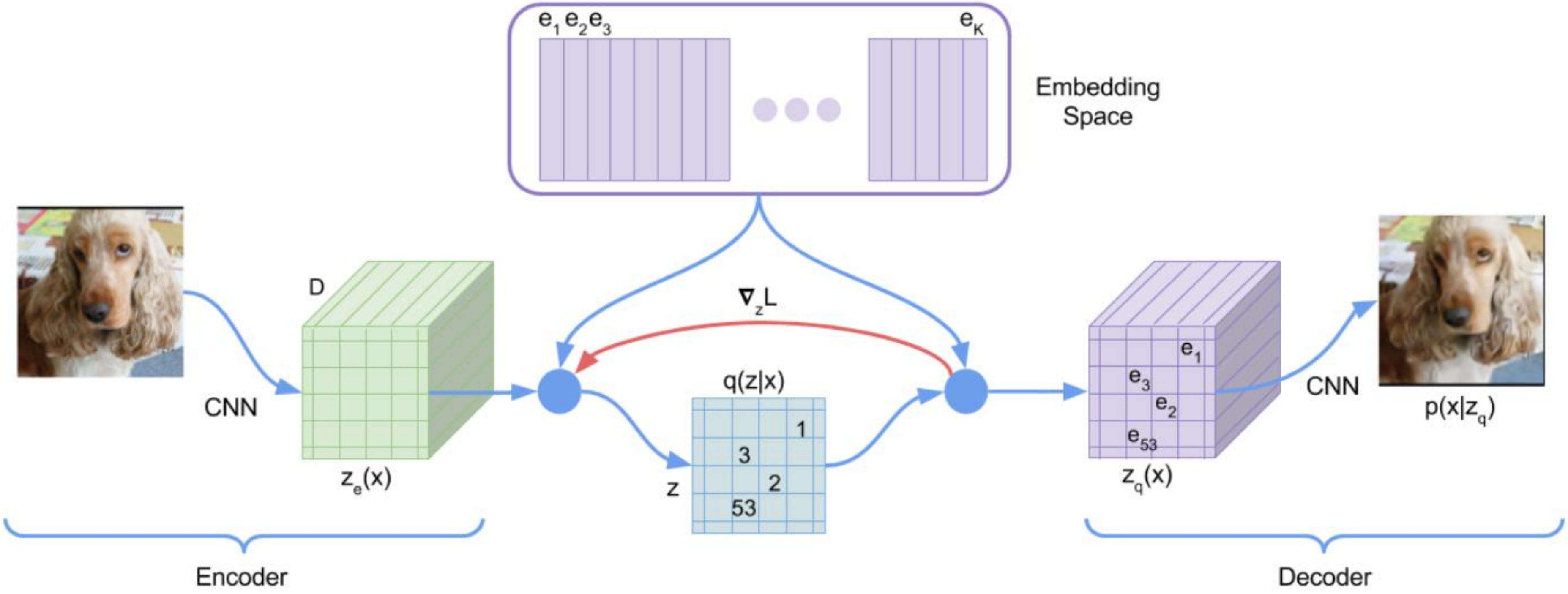
Jukebox: A Generative Model for Music

Speaker: Petr Grinberg

Recall: VQ-VAE



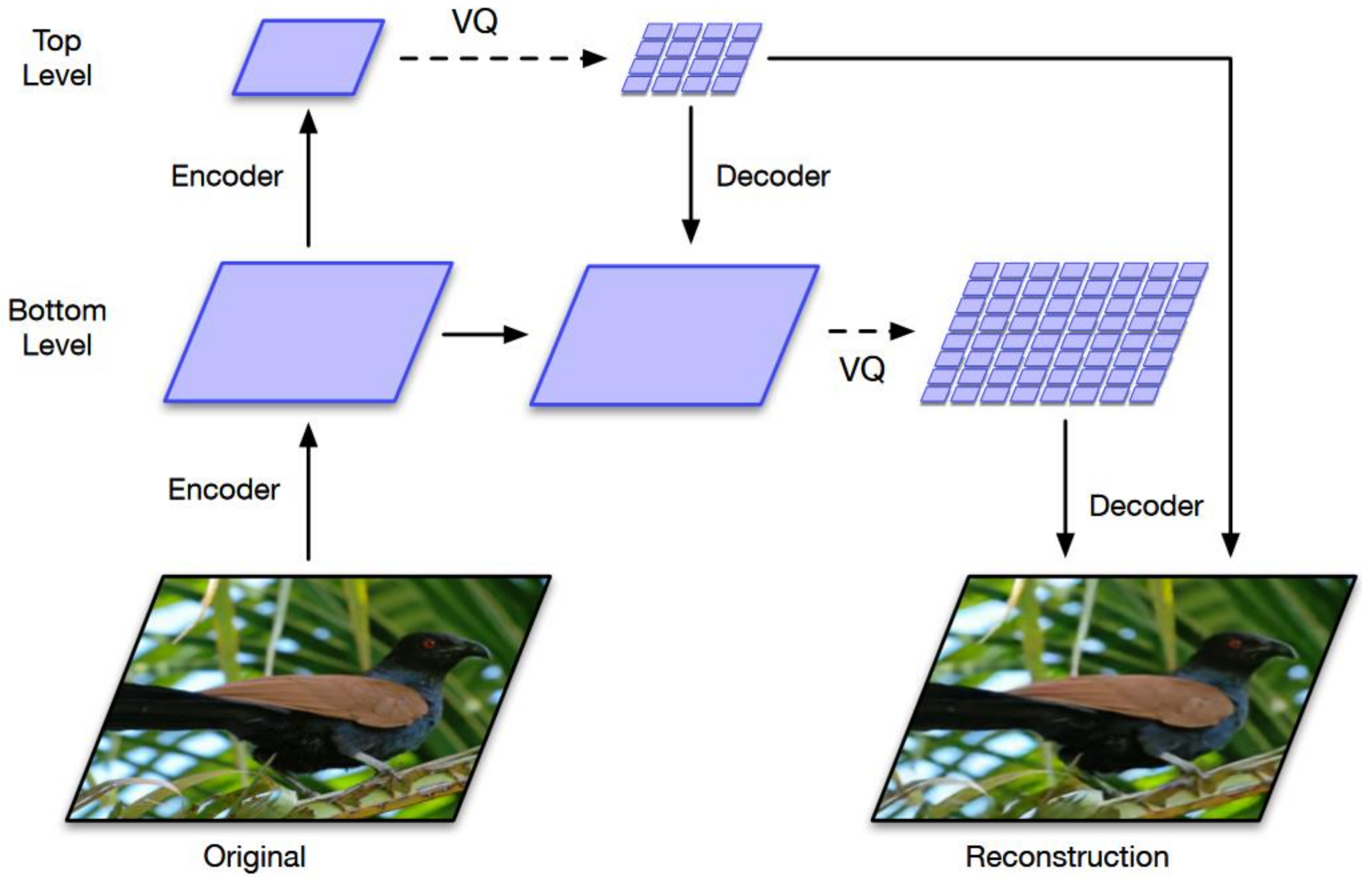
New pictures can be generated from autoregressive distribution over z, p(z)

Recall: VQ-VAE, Training Loss

$$
\begin{aligned}
\mathcal{L} &= \mathcal{L}_{\text{recons}} + \mathcal{L}_{\text{codebook}} + \beta\mathcal{L}_{\text{commit}} \\
\mathcal{L}_{\text{recons}} &= \tfrac{1}{T}\sum_{t} \|\mathbf{x}_t - D(\mathbf{e}_{z_t})\|_2^2 \\
\mathcal{L}_{\text{codebook}} &= \tfrac{1}{S}\sum_{s} \|\mathrm{sg}\left[\mathbf{h}_s\right] - \mathbf{e}_{z_s}\|_2^2 \\
\mathcal{L}_{\text{commit}} &= \tfrac{1}{S}\sum_{s} \|\mathbf{h}_s - \mathrm{sg}\left[\mathbf{e}_{z_s}\right]\|_2^2
\end{aligned}
$$

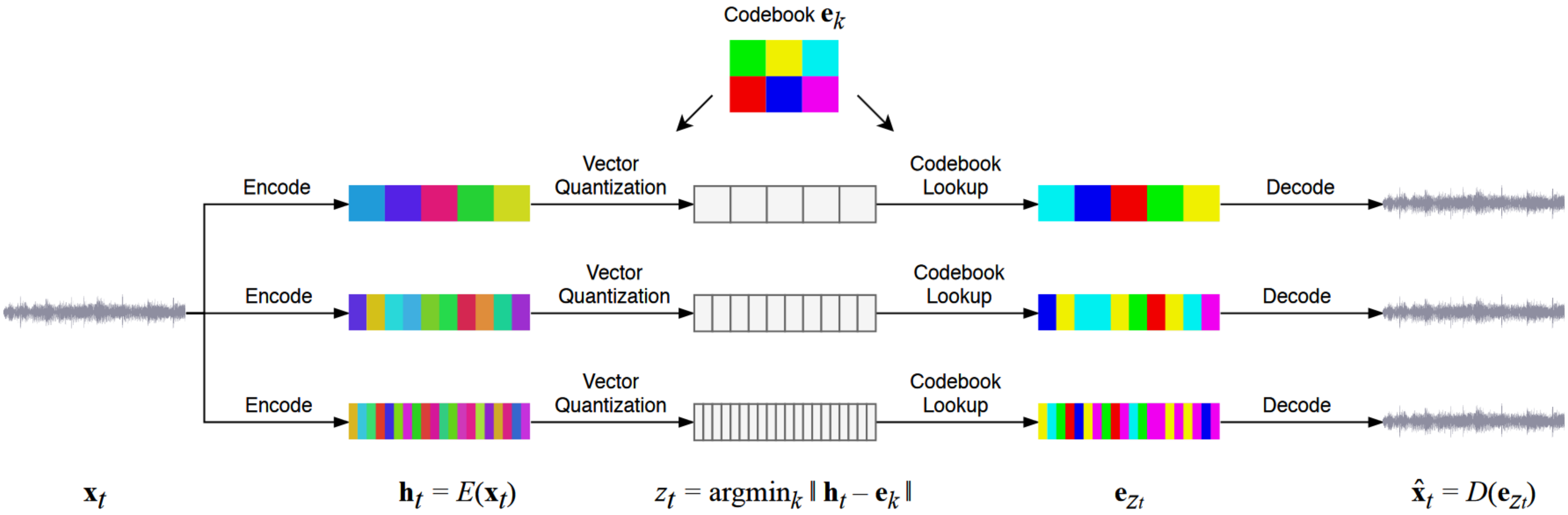
Reconstruction Loss can vary, this one will be used in Jukebox

Recall: VQ-VAE2 (Hierarchical)

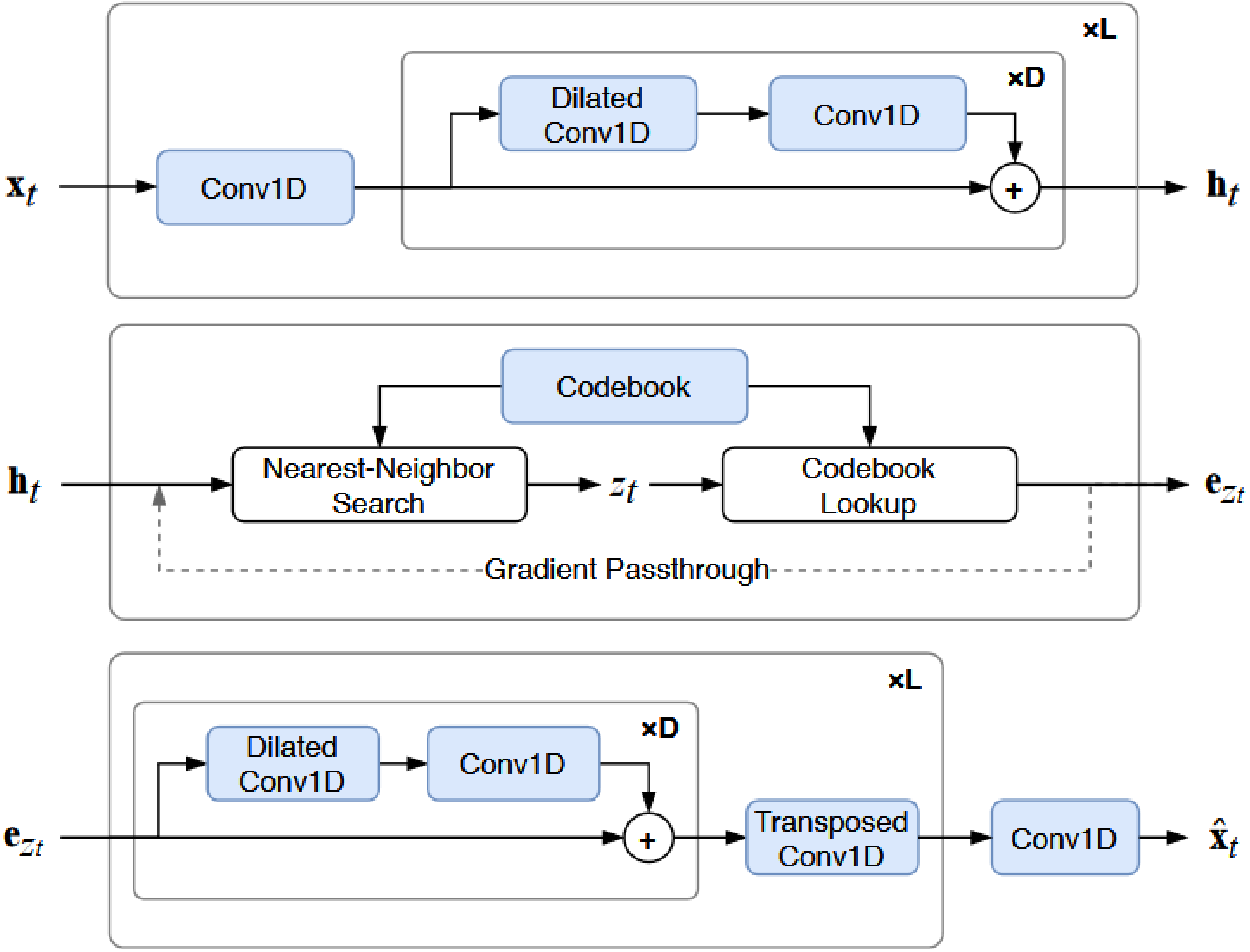


Bottom Level: local features, Top Level: high-level semantics

Jukebox: VQ-VAE



Jukebox: VQ-VAE



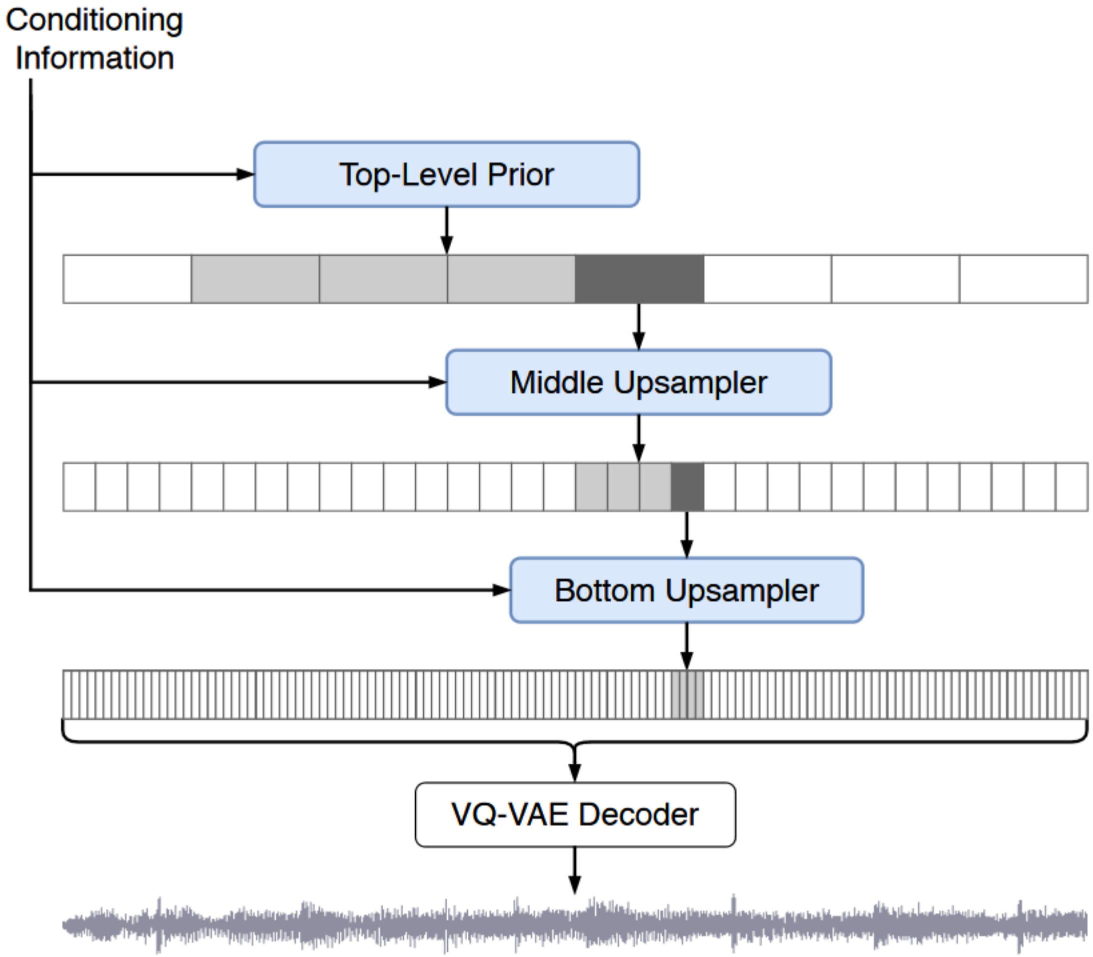
Jukebox: VQ-VAE, Training Hacks

- Random restarts for embeddings + EMA
- Separated Autoencoders
- Spectral Loss (for high-frequencies reconstruction):

$$\mathcal{L}_{\text{spec}} = \| |\text{STFT}(\mathbf{x})| - |\text{STFT}(\widehat{\mathbf{x}})| \|_2$$

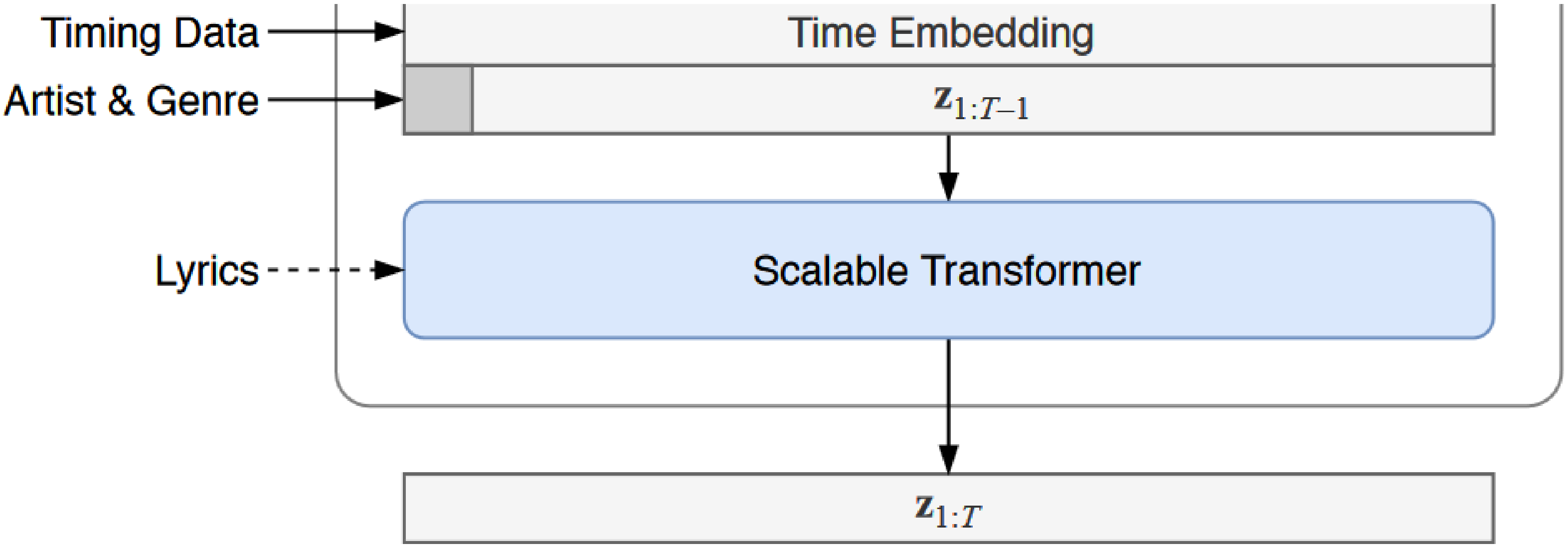
Jukebox: Learning Prior

$$\begin{aligned} p(\mathbf{z}) &= p(\mathbf{z}^{\text{top}}, \mathbf{z}^{\text{middle}}, \mathbf{z}^{\text{bottom}}) \\ &= p(\mathbf{z}^{\text{top}})p(\mathbf{z}^{\text{middle}}|\mathbf{z}^{\text{top}})p(\mathbf{z}^{\text{bottom}}|\mathbf{z}^{\text{middle}}, \mathbf{z}^{\text{top}}) \end{aligned}$$

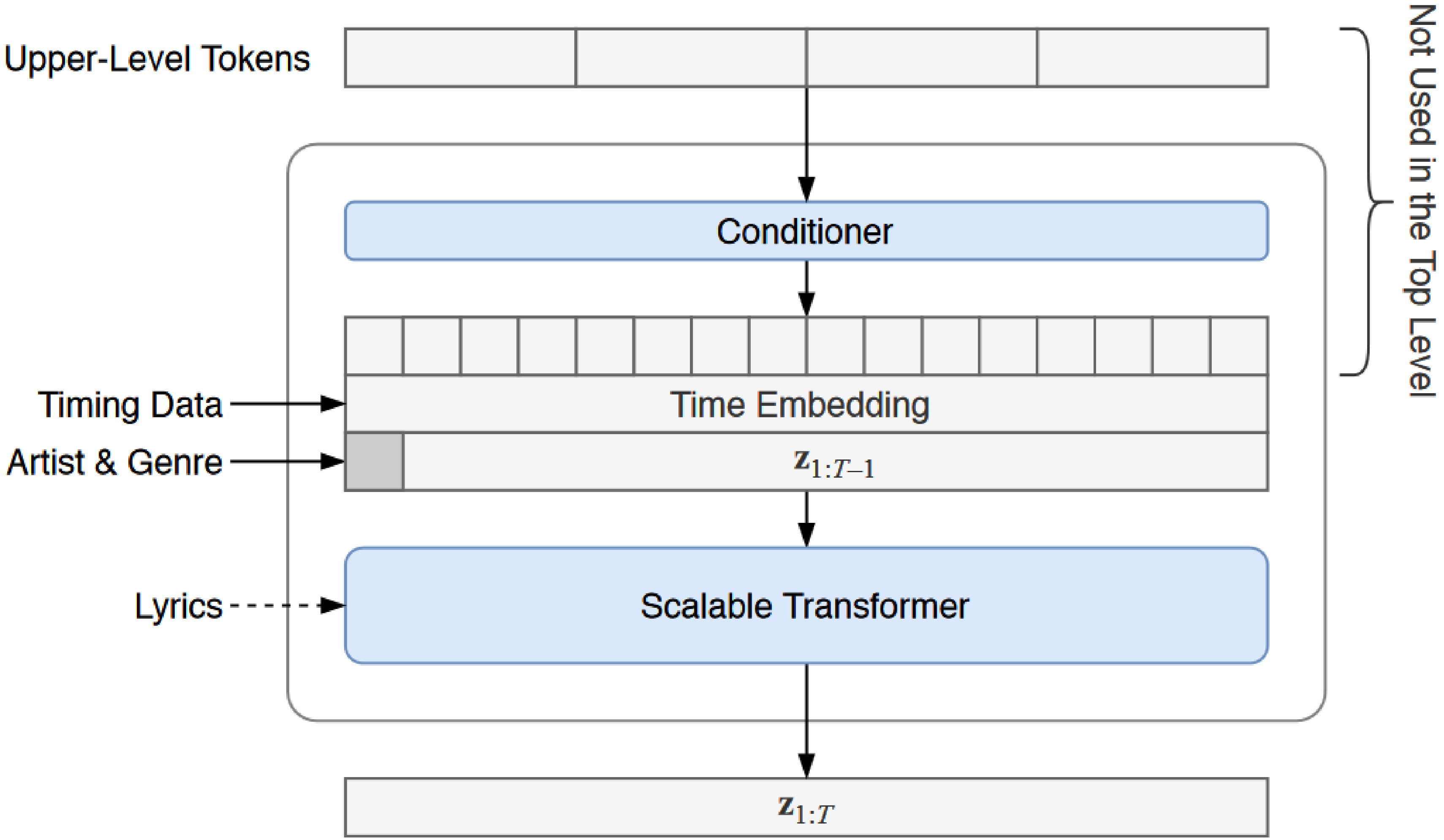


Jukebox: Learning Prior

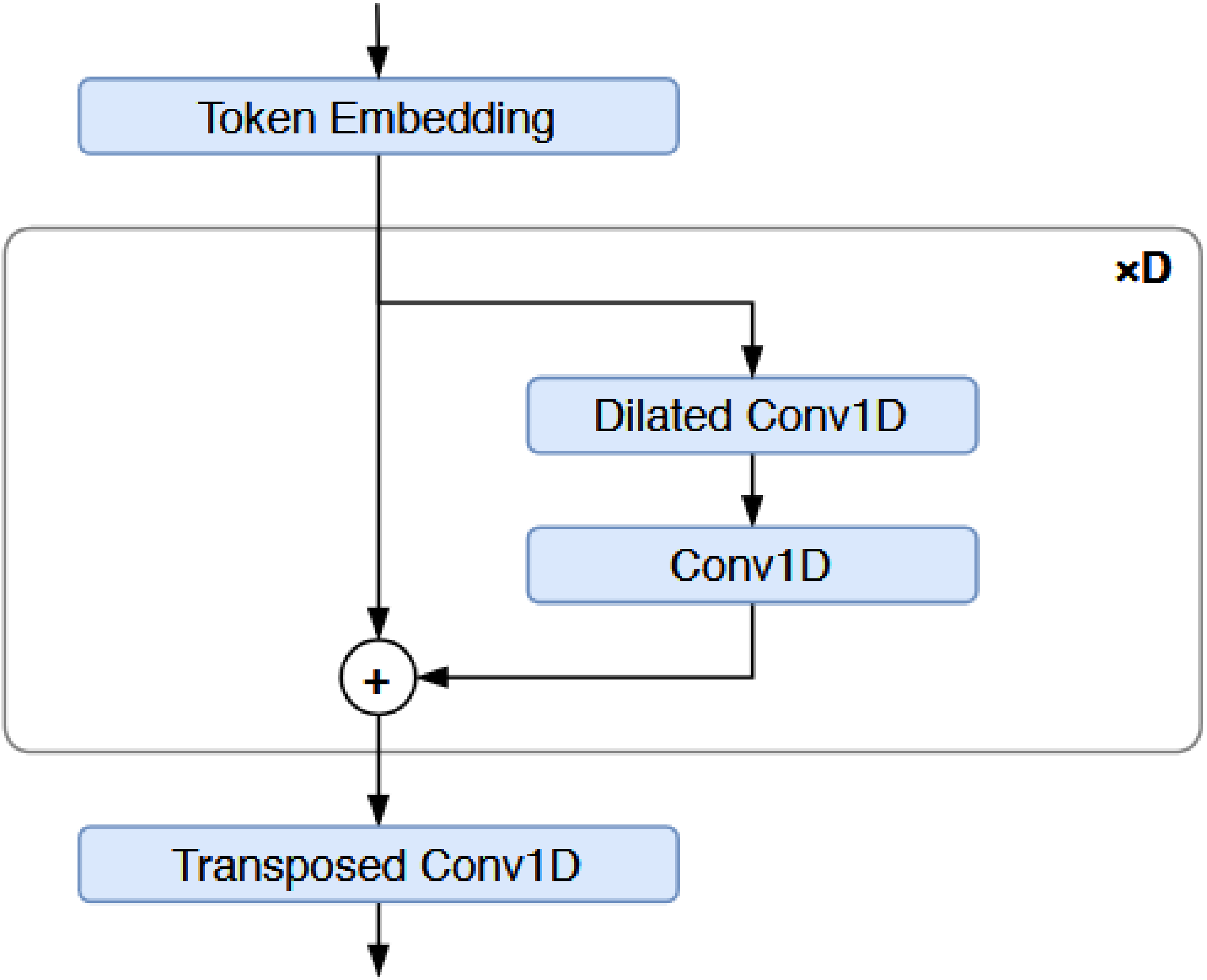
Top-level prior architecture:



Jukebox: Learning Prior

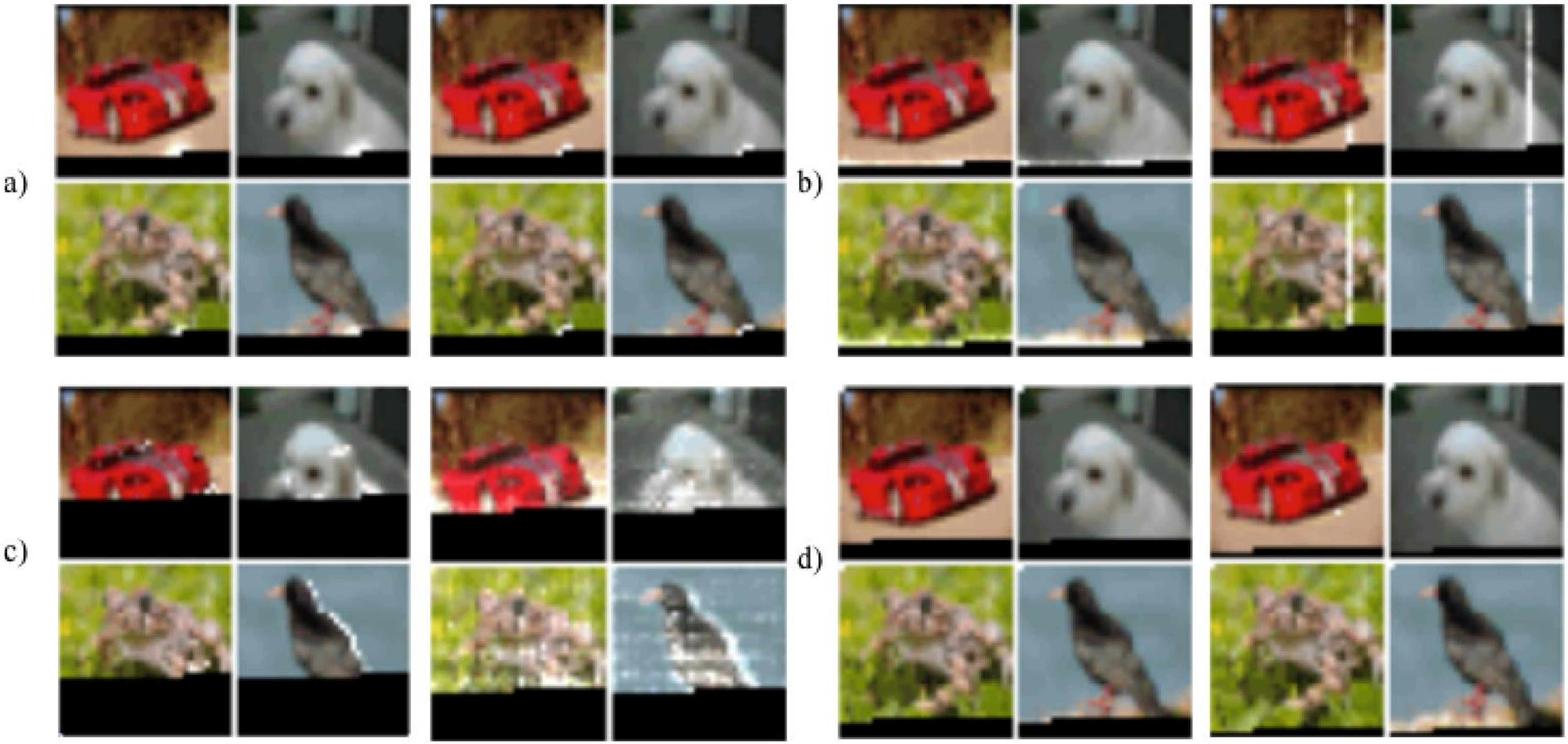


Jukebox: Conditioner



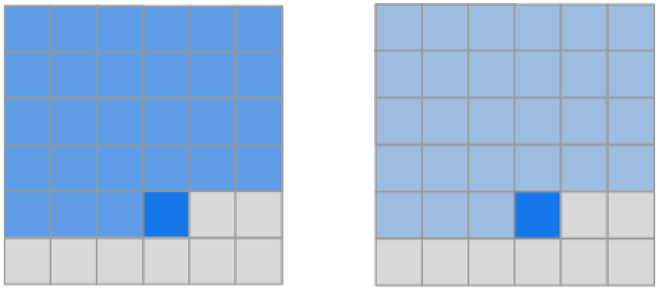
Jukebox: Scalable Transformer

Firstly, let’s recall Sparse Transformer:

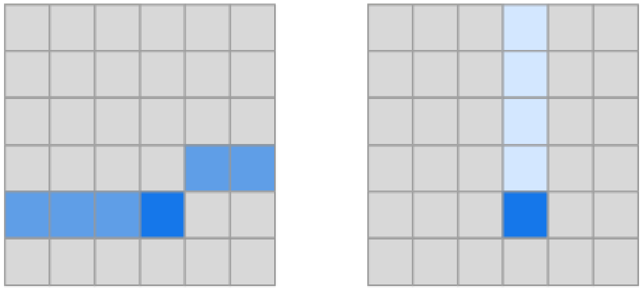


Jukebox: Scalable Transformer

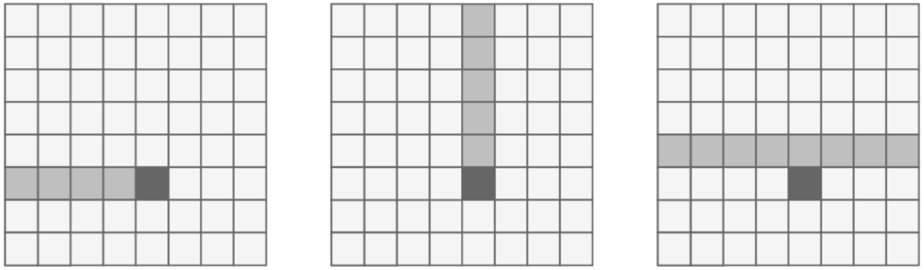
- Vanilla Transformer



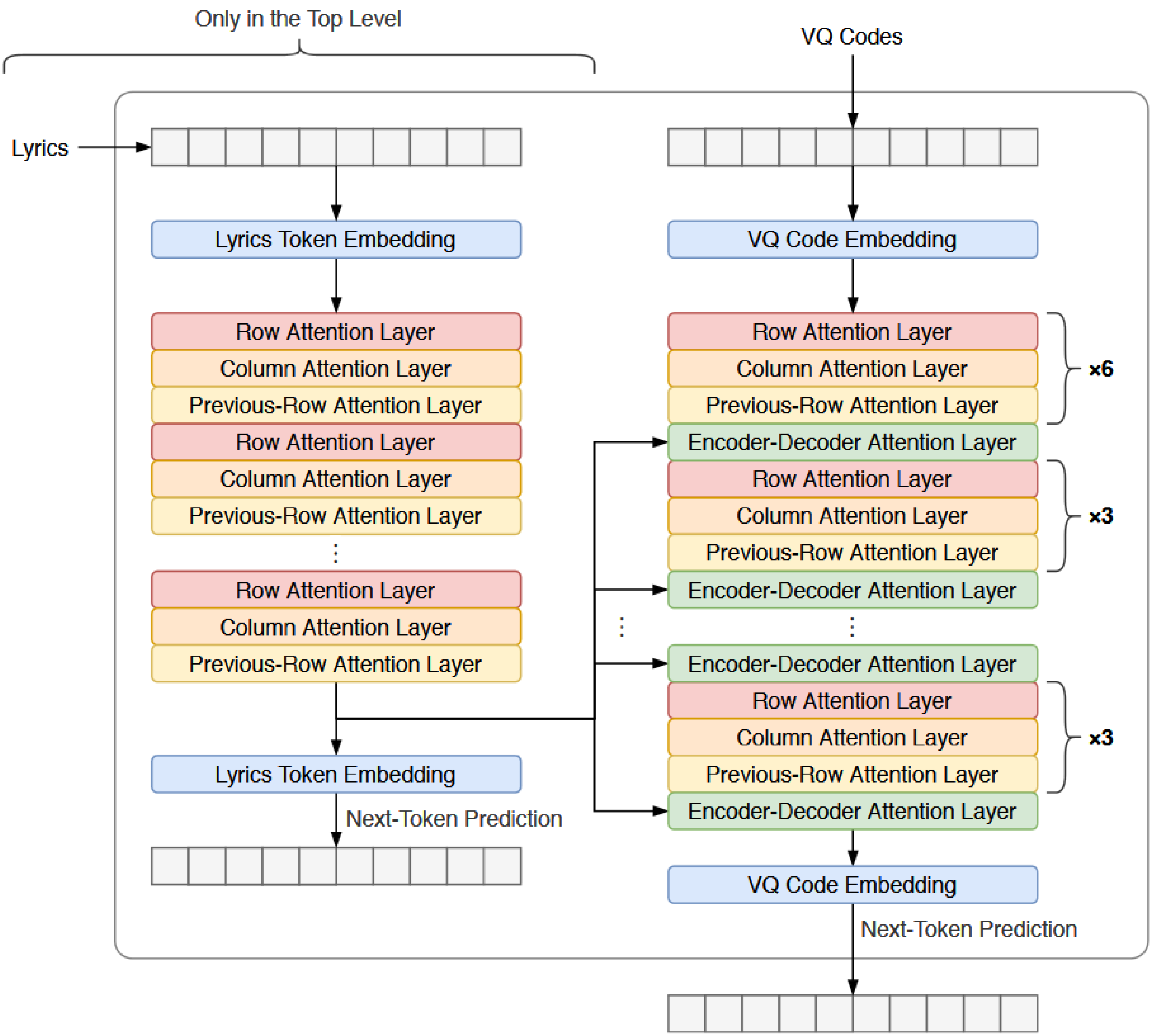
- Sparse Transformer



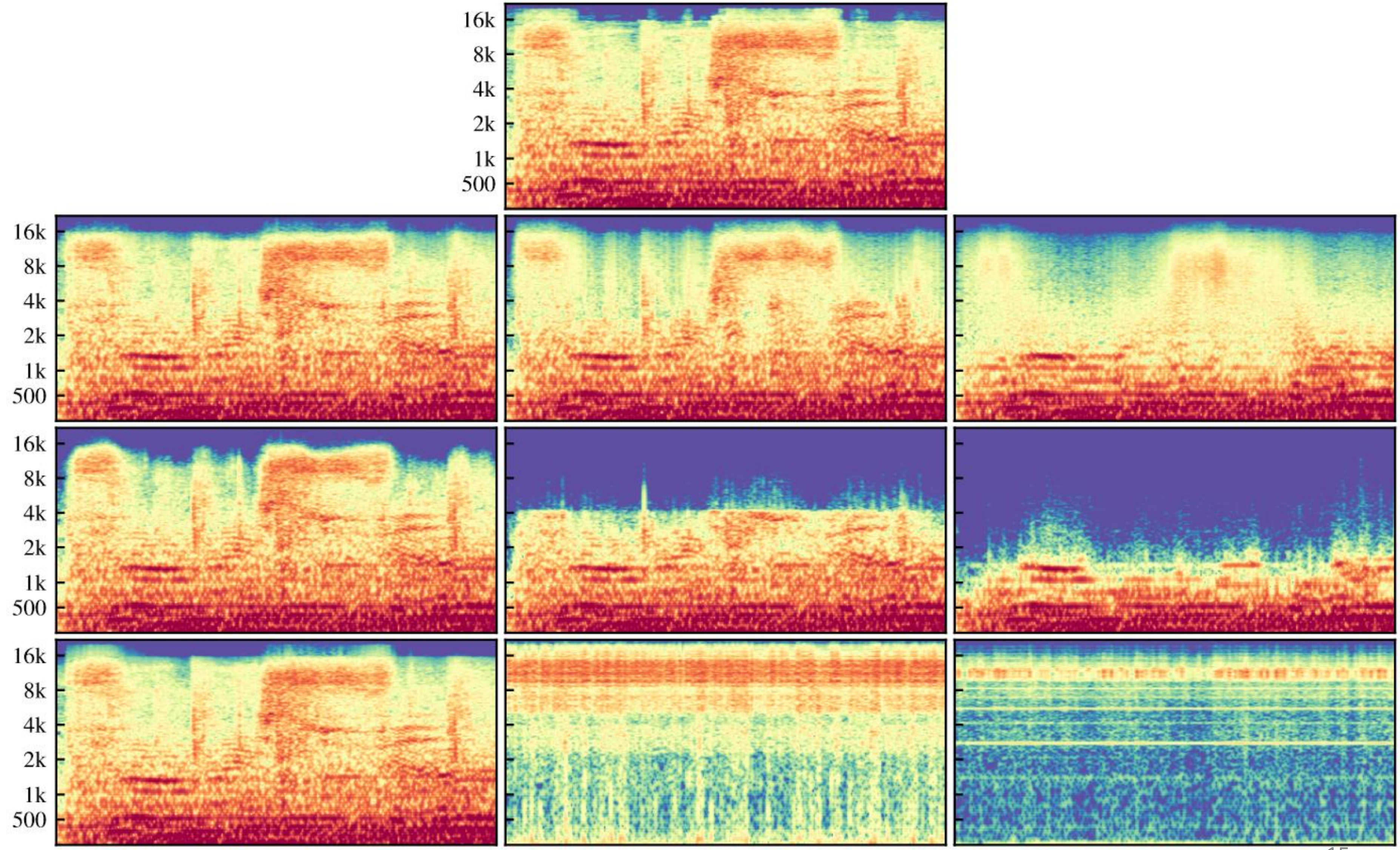
- Scalable Transformer



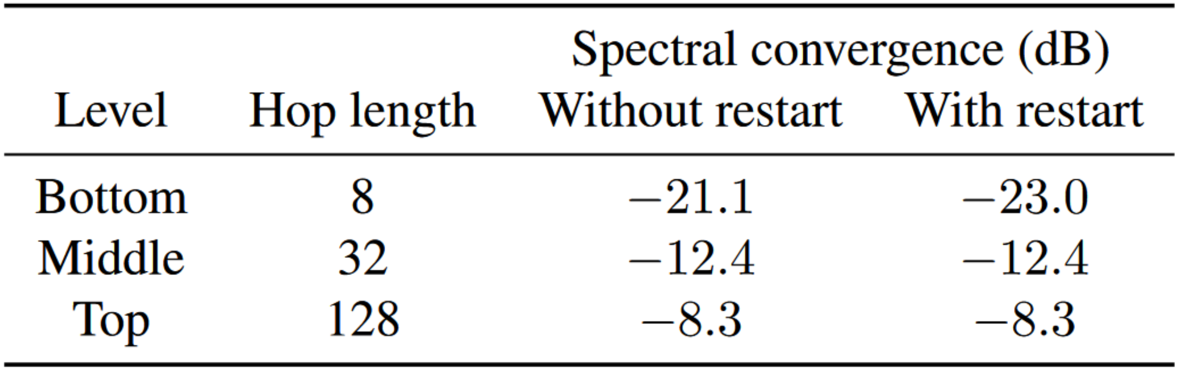
Jukebox: Lyrics Conditioning



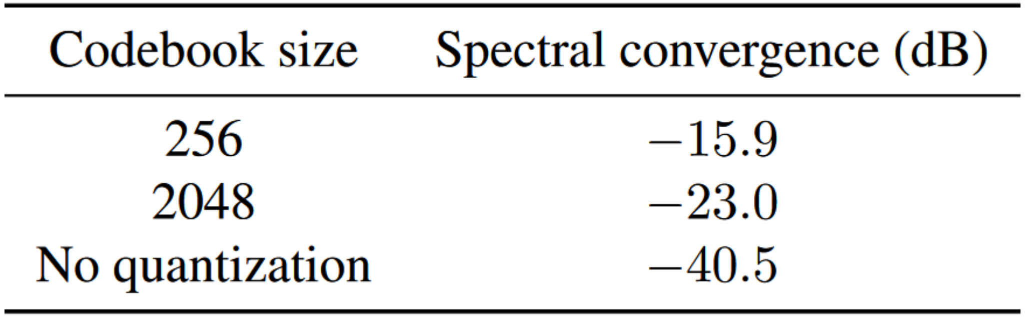
Jukebox: Ablation Study



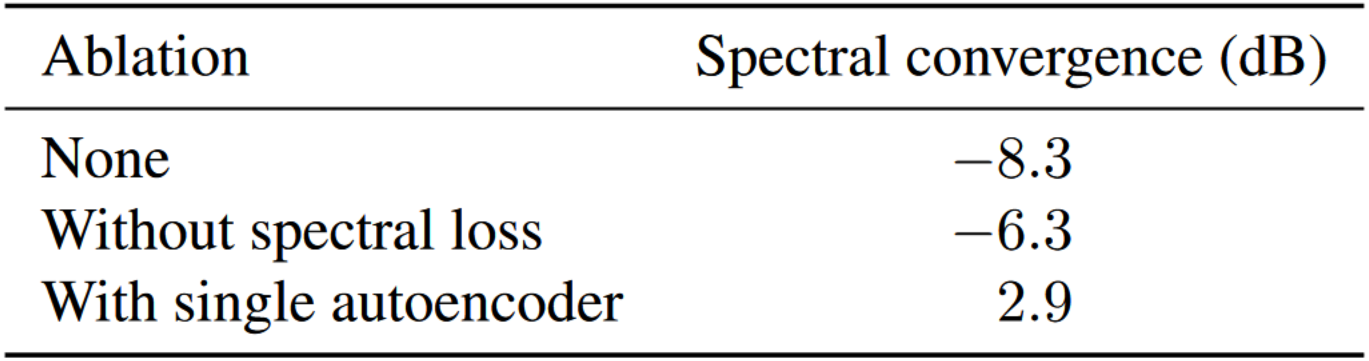
Jukebox: Ablation Study



		Spectral convergence (dB)	
Level	Hop length	Without restart	With restart
Bottom	8	−21.1	−23.0
Middle	32	−12.4	−12.4
Top	128	−8.3	−8.3



Codebook size	Spectral convergence (dB)
256	−15.9
2048	−23.0
No quantization	−40.5



Ablation	Spectral convergence (dB)
None	−8.3
Without spectral loss	−6.3
With single autoencoder	2.9