Perceptually Aligning Representations of Music via Noise-Augmented Autoencoders

Mathias Rose Bjare* **Giorgia Cantisani†** **Marco Pasini‡**
Stefan Lattner§ **Gerhard Widmer***
* Johannes Kepler University, Linz, AT, †ENS, PSL University, CNRS, Paris, FR,
‡Queen Mary University, London, UK, § Sony Computer Science Laboratories (CSL), Paris, FR
mathias.bjare@jku.at

Abstract

We argue that training autoencoders to reconstruct inputs from noised versions of their encodings, when combined with perceptual losses, yields encodings that are structured according to a perceptual hierarchy. We demonstrate the emergence of this hierarchical structure by showing that, after training an audio autoencoder in this manner, perceptually salient information is captured in coarser representation structures than with conventional training. Furthermore, we show that such perceptual hierarchies improve latent diffusion decoding in the context of estimating surprisal in music pitches and predicting EEG-brain responses to music listening. Pretrained weights are available on `github.com/CPJKU/pa-audioic`.

1 Introduction

Essential aspects of music appreciation, composition, and cognition are musical self-similarity, which sets expectations about the continuation of the music being listened to, and the consequent novelty or *surprisal* arising as the incoming sensory input confronts these expectations. The computational estimation of perceived musical expectations and surprisal has been studied using *information content* (IC) or negative log-likelihood (NLL) of autoregressive models [1–5]. The correlation between IC and surprisal has been perceptually validated in numerous behavioral and neural studies [3, 5–10]. Due to challenges calculating IC, previous research has mainly focused on the monophonic symbolic music data [3, 8–10]. In the audio domain, [11, 12] has proposed estimating musical surprise using Bayesian predictive inference on sequences of audio features. However, both approaches are limited to a few hand-selected music features that ignore much of the audio signal stimuli used in listener experiments. To overcome this and the high dimensionality of audio, [5] estimates musical surprisal from audio autoencoder latent representations using the IC of autoregressive models. The methodology has recently been extended to more powerful autoregressive diffusion models in [13]. Notably, IC can be computed at different stages of the diffusion process, which correspond to varying levels of "noise" in the data. The authors show that for appropriately moderate noise levels, the suprisal of important musical features, such as pitch, is better estimated. The authors hypothesize that at these noise levels, most pitch-related information is present, while information of less relevance to pitch, such as timbre nuances, is less dominant. Spectral analysis of ordinary diffusion forward processes reveals that all frequencies of the signal entering the process (in our case, autoencoder representations) are noised equally with a strength that increases with higher noise levels [14, 15]. As a result, low spectral power structures (fine structures) of the signal are indistinguishable from the noise in the mixed signal and, therefore, provide no gradient to the denosing network, at lower noise levels than structures with high spectral power (coarse structures). In the following, we refer to this as the spectral signal-to-noise ratio (SNR) properties of diffusion noise processes. The underlying hypothesis of [13] is that an alignment between coarse structures in representations and perceptual features (such as pitch-like qualities) exists. However, this is typically not enforced explicitly during autoencoder training.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

In this paper, we show that a recent autoencoder training technique [16], which adds varying amount of noise to the latents during training, when combined with traditional perceptual loss objectives, hierarchically aligns perceptual features with latent structure — such that the most salient perceptual information is captured in the coarsest structures, while progressively finer structures encode less perceptually relevant information. Furthermore, aligning coarser structures with more important perceptual information might increase diffusion decoding performance in general, as diffusion models produce more accurate denoisings for coarser structures than finer structures. This is due to the inverted U-shaped properties of the loss and modern diffusion noise-schedules [17, 18]. See Appendix A for related work on perceptual alignment in the image-pixel domain. We demonstrate the learning of perceptual hierarchies by finetuning the Music2Latent [19] autoencoder with noise-augmented latents and show that reconstructions from latents with varying amounts of noise preserve perceptual information better in aligned latent spaces than in unaligned spaces. Furthermore, we demonstrate the importance of perceptual latent alignment for latent diffusion decoding in the case of musical surprisal estimation. Specifically, we train autoregressive diffusion models in the aligned space to estimate surprisal in vocal and synthetic music. Our results show that surprisal estimation is improved by the alignment procedure, as demonstrated by higher correlations with predictions of a rigorously perceptually validated [8–10] symbolic pitch expectancy model and in terms of predicting EEG brain responses to vocal music. The estimation, furthermore, improves on the results of previous methods. Moreover, we find the best estimations in aligned latent spaces at intermediate noise levels, whereas in unaligned spaces this is not always the case. This further supports that the aligned representations contain more important perceptual information in coarse structures than unaligned representations.

2 Latent diffusion

Latent diffusion consists of two stages. Firstly, an autoencoder is trained to produce highly compressed data representations. Secondly, a diffusion model is trained to reproduce latent encoded data. For the first stage, we employ the consistency autoencoder (CAE) of [19], composed of the encoder–decoder pair (E, D). Given an input audio sample x, the encoder produces a compressed latent representation $z = E(x)$, and the decoder reconstructs the signal as $\hat{x} = D(z)$. In the CAE, D is a (stochastic) consistency model [20] that is conditioned on the outputs of E. The model is trained via a consistency training [20], which implicitly minimizes a perceptually weighted [21] complex spectrogram difference between reconstruction and input. In fact, modern autoencoders for latent diffusion like [22] typically include some perceptual loss, either in the reconstruction loss or as an additional loss [16]. For the second stage, we train an autoregressive rectified flow model [13, 23, 24]: a rectified flow model [25, 26] to generate next-step predictions conditioned on a context embedding of past observations summarized by a transformer [27]. In this paper, instead of generating samples autoregressively, we compute IC or negative log-likelihoods of next-step predictions in a teacher-forcing manner using the instantaneous change of variables formulae [28] as in [13].

3 Noised reconstruction training and perceptual alignment

[16] studies noise-augmenting the traditional autoencoder reconstruction learning framework for diffusion models by interpolating the latents with noise similar to the noising process of rectified flows. During training, [16] noises latents z with

$$z' = (1-t)z + t\,n(\gamma), \quad \text{where} \quad n(\gamma) \sim \gamma \cdot \mathcal{N}(0, I),\ t \sim \mathcal{U}(0,1). \tag{1}$$

and task the autoencoder to reconstruct clean data. Although noising the latents during autoencoder training seems similar to diffusion forward noise processes, we argue that it serves a fundamentally different purpose since z is learned and not frozen. Observing the reconstruction of a single input data example, the encoder has to learn representations that, when decoded, simultaneously minimize the perceptual loss for different noise levels. Following the spectral SNR properties of diffusion noise processes, this particularly means that information related to satisfying the perceptual loss should mostly be encoded in coarse latent structures, and information with increasingly less perceptual relevance should be encoded in increasingly finer structures.

We propose the following modifications to the method presented in [16] that we empirically found beneficial for downstream tasks. Using the noise process of Equation (1), the expected SNR is given

Table 1: Perceptual quality metrics for reconstructions of aligned latents $NT = E, D$ and unaligned latents $NT = \emptyset$ and $NT = D$.

NT	SNR	V (↑)	SI (↑)	F_{VGG} (↓)	F_{CLAP} (↓)
E, D	∞	3.73	-5.18	1.53	0.05
	4.0	3.48	-9.05	2.46	0.08
	1.0	3.19	-15.73	3.64	0.17
	0.25	2.87	-29.78	5.01	0.38
D	∞	3.73	-4.97	1.58	0.05
	4.0	3.45	-10.31	2.89	0.09
	1.0	3.18	-18.52	3.94	0.19
	0.25	2.88	-32.17	5.10	0.35
$\emptyset$	∞	3.84	-3.84	1.16	0.04
	4.0	2.94	-11.44	6.63	0.42
	1.0	2.53	-18.82	11.15	0.84
	0.25	2.22	-28.44	15.04	1.17

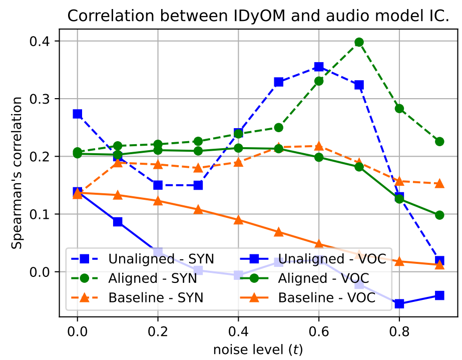

Figure 1: Correlation between IC calculated on melodies using IDyOM and calculated with aligned and unaligned latents and using the baseline of [13] at different noise levels.

by $\mathbb{E}[z^2]/\gamma^2$, and can be controlled by γ. The encoder in [16], however, can learn to increase the variance of z to increase the expected SNR, which essentially reduces the effect of noising. We fix the variance of z to the variance of the noise distribution using layer normalization [29], such that the expected SNR stays constant during training. We set $\gamma^2 = 1$ and control the expected SNR by sampling t from a biased logit-normal [30] distribution sigmoid(ε), where $\varepsilon \sim \mathcal{N}(m, s^2)$ as in [18]. Unlike in [16], our latent noise process is the same as the rectified flow noise process used for latent diffusion, except for t's distribution. In Appendix B, we show the importance of fixing the variance and provide model selection details.

4 Reconstruction experiments

We test the efficiency of the noising technique in hierarchically aligning more perceptually important features with coarse structures using the autoencoders' reconstructions. We finetune the publicly available Music2Latent checkpoint using the same data, architecture, and hyperparameters as described in [19], except that we use a constant consistency-step schedule with a step size fixed to the final value of the pre-trained model. To quantify the amount of perceptually important features encoded in the latents, we decode them, and check if they perceptually correspond to the input reference, using the reconstruction metrics ViSQOL (V) [31–33], a MOS-like distance between two audio samples, and SI-SDR (SI), a spectrogram distance [34], as used in [19, 35]. Since it is challenging to disentangle structures of varying coarseness in the latents explicitly, we instead use the spectral SNR-properties and construct latents at four different coarseness levels by encoding diverse 10s music audioclips from MusicCaps [36] and adding noise following the same latent noising process as used for training, but setting t in a way that the SNR levels are $\{\infty, 4, 1, .25\}$ respectively. We report the results in Table 1 as $NT = E, D$, indicating that both encoder and decoder have been trained with noised latents. At low SNR levels, it is likely that information essential to faithfully reconstructing the signal is removed, leaving the decoder to infer the likely form of the input. To measure the realism of reconstructions at low SNR levels (not necessarily following the input strictly), we additionally report the distribution metrics FAD [37] score using the VGGISH (F_{VGG}) and the CLAP (F_{CLAP})[38] versions. We compare the results of the perceptually aligned autoencoder against the results of an unaligned autoencoder in two scenarios: 1) where the training-inference discrepancy is fixed by freezing the encoder and training the decoder with noised latents, reported as $NT = D$, and 2) without the correction, reported as $NT = \emptyset$. Comparing SI for $NT = E, D$ and $NT = D$ at SNR values $< \infty$, we find that the perceptual information retained in coarse structures is always higher for the aligned representations than for the unaligned representations and similar for V. Both V,SI are higher comparing $NT = E, D$ and $NT = \emptyset$ for $SNR < \infty$, except for SI at $SNR = 0.25$. At this low SNR level, where much of the input's information has been removed, the FAD score of the aligned autoencoder is much lower. This indicates that the stochastic decoder is inventing information to create plausible reconstructions that diverge from the input.

5 Musical surprisal estimation experiments

We are interested in whether hierarchical latent alignment improves musical surprisal estimation in the diffusion noise space continuum. Specifically, we investigate the model's capabilities to estimate pitch surprisal and to predict EEG responses to sung music. For estimating pitch surprisal, we largely follow the methodology of [13] and train an autoregressive rectified flow model using the same data, model, and hyperparameters except for the changes mentioned in Appendix B.2. We then compare whether alignment improves agreement between IC derived from autoregressive latent diffusion models and IC derived from IDyOM [3]; a perceptually validated [6–10] pitch expectancy model that operates in a condensed symbolic domain. We conduct our experiment using a synthesized dataset (SYN) of Irish monophonic tunes, described in [39], and a recorded vocal dataset (VOC) along with its automatic transcription, described in [40]. We extract the IC of each note pitch in the symbolic datasets using IDyOM and pair these with IC values calculated with our models in aligned and unaligned spaces at various noise levels. We compare the paired estimates using Spearman's rank correlation and report the results in Figure 1, which are significant on a 5% significance level (except for VOC unaligned correlations $t \in [0.3, 0.7]$). We additionally report the results of [13] as Baseline. For the aligned latent space, as the noise level decreases, the correlation increases until a maximum and then decreases. This suggests that after reaching a certain noise level, within $[0.5, 0.7]$ for both datasets, most information relevant for pitch surprisal estimation is already present in the noised data, and adding additional information may decrease the correlation. For the unaligned space and the baseline, this is not the case. The highest correlations for both datasets are found in the aligned space.

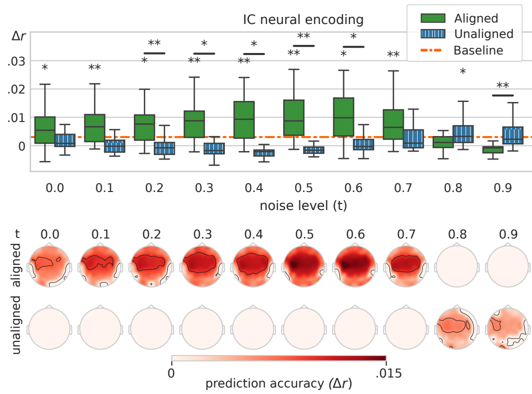

Figure 2: Cortical tracking of IC computed with aligned and unaligned latents across different noise levels. Δr denotes the increase in prediction accuracy when comparing a full model (IC + acoustic envelope) with a reduced model including only the envelope. Bar plots report the mean ± SE across participants (median across electrodes, average across trials). Scalp topographies report Δr for individual channels (only significant channels are shown, significance threshold at $p < 0.05$).

To further evaluate how the proposed model correlates with human perception, we tested whether IC estimated with perceptually aligned latents predicts neural responses to music more accurately than unaligned. To do so, we compared the neural encoding of IC features computed with aligned and unaligned latents across different noise levels in EEG responses to the sung music of VOC (64 channels, 20 participants, 18 songs, see Appendix B.3 and [40] for details) and report the results in Figure 2. The IC of the aligned method produced significantly stronger cortical tracking than the IC of the unaligned method and the baseline model of [13] across most noise levels ($t = 0.2$–0.6), with the largest improvements observed around mid-level noise. This is consistent with the highest IDyOM correlations observed in that dataset. This advantage was consistent across participants and electrodes, as also reflected in the scalp topographies, which revealed widespread positive effects over fronto-central regions. Taken together, these findings demonstrate that (i) perceptually aligned IC is reliably encoded in neural responses to music, (ii) but only under moderate noise level conditions.

Together, the two musical surprisal experiments yield consistent results, suggesting that perceptually structured latent representations may benefit tasks in music perception and latent-diffusion decoding. This is despite the information loss caused by the noise augmentation, as suggested by the lower reconstructive performance on clean data, and is similar to the findings of [16]. Future studies should investigate whether closing the gap leads to better overall results. In our EEG experiments, we obtained the best results at a high noise level ($t = 0.6$). Interestingly, this noise level coincides with a high performance in our pitch suprisal estimation and is consistent with the result of [41] that predicts EEG responses only using pitch information. Future work should investigate other musical or audio features present in the signal at that noise level, as well as identify those that emerge at lower noise levels, since their presence appears to impair EEG prediction.

6 Acknowledgments

The work leading to these results was conducted in a collaboration between JKU and Sony Computer Science Laboratories Paris under a research agreement. The first and fifth author also acknowledge support by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement 101019375 (*"Whither Music?"*).

References

[1] Leonard B Meyer. Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, 15(4):412–424, 1957.

[2] Darrell Conklin and Ian H Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.

[3] Marcus Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computing, City University, London, UK, 2005.

[4] Mathias Rose Bjare, Stefan Lattner, and Gerhard Widmer. Controlling surprisal in music generation via information content curve matching. In *ISMIR*, 2024.

[5] Mathias Rose Bjare, Giorgia Cantisani, Stefan Lattner, and Gerhard Widmer. Estimating musical surprisal in audio. In *ICASSP*, 2025.

[6] Marcus T Pearce, María Herrojo Ruiz, Selina Kapasi, Geraint A Wiggins, and Joydeep Bhattacharya. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, 50(1):302–313, 2010.

[7] Giovanni M Di Liberto, Claire Pelofi, Roberta Bianco, Prachi Patel, Ashesh D Mehta, Jose L Herrero, Alain De Cheveigné, Shihab Shamma, and Nima Mesgarani. Cortical encoding of melodic expectations in human temporal cortex. *Elife*, 9:e51784, 2020.

[8] Niels Chr Hansen and Marcus T Pearce. Predictive uncertainty in auditory sequence processing. *Frontiers in psychology*, 5:1052, 2014.

[9] Roberta Bianco, Lena Esther Ptasczynski, and Diana Omigie. Pupil responses to pitch deviants reflect predictability of melodic sequences. *Brain and Cognition*, 138:103621, 2020.

[10] Toviah Moldwin, Odelia Schwartz, and Elyse S Sussman. Statistical learning of melodic patterns influences the brain's response to wrong notes. *Journal of cognitive neuroscience*, 29 (12):2114–2122, 2017.

[11] Benjamin Skerritt-Davis and Mounya Elhilali. Detecting change in stochastic sound sequences. *PLoS Comput. Biol.*, 14(5), 2018.

[12] Benjamin Skerritt-Davis and Mounya Elhilali. A model for statistical regularity extraction from dynamic sounds. *Acta Acustica united with Acustica*, 105(1):1–4, 2019.

[13] Mathias Rose Bjare, Stefan Lattner, and Gerhard Widmer. Estimating musical surprisal from audio in autoregressive diffusion model noise spaces. In *ISMIR*, 2025.

[14] Sander Dieleman. Diffusion is spectral autoregression, 2024. URL `https://sander.ai/2024/09/02/spectral-autoregression.html`.

[15] Fabian Falck, Teodora Pandeva, Kiarash Zahirnia, Rachel Lawrence, Richard E. Turner, Edward Meeds, Javier Zazo, and Sushrut Karmalkar. A fourier space perspective on diffusion models. *CoRR*, abs/2505.11278, 2025.

[16] Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good visual tokenizers. *arXiv preprint arXiv:2507.15856*, 2025.

[17] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

[18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*. OpenReview.net, 2024.

[19] Marco Pasini, Stefan Lattner, and George Fazekas. Music2latent: Consistency autoencoders for latent audio compression. In *ISMIR*, 2024.

[20] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, volume 202, pages 32211–32252, 2023.

[21] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2351–2364, 2023.

[22] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP*, pages 1–5. IEEE, 2025.

[23] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024.

[24] Marco Pasini, Javier Nistal, Stefan Lattner, and George Fazekas. Continuous autoregressive models with noise augmentation avoid error accumulation. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[25] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

[26] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*. OpenReview.net, 2023.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[28] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *NeurIPS*, pages 6572–6583, 2018.

[29] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[30] Jhon Atchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

[31] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

[32] Colm Sloan, Naomi Harte, Damien Kelly, Anil C Kokaram, and Andrew Hines. Objective assessment of perceptual audio quality using visqolaudio. *IEEE Transactions on Broadcasting*, 63(4):693–705, 2017.

[33] Michael Chinen, Felicia SC Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. Visqol v3: An open source production ready objective speech and audio metric. In *2020 twelfth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2020.

[34] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[35] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[36] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[37] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, pages 2350–2354. ISCA, 2019.

[38] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pages 1–5. IEEE, 2023.

[39] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova. Music transcription modelling and composition using deep learning. In *Proceedings of the Conference on Computer Simulation of Musical Creativity*, Huddersfield,UK, 2016.

[40] Giorgia Cantisani, Amirhossein Chalehchaleh, Giovanni Di Liberto, and Shihab Shamma. Investigating the cortical tracking of speech and music with sung speech. In *INTERSPEECH*, pages 5157–5161. ISCA, 2023.

[41] Giorgia Cantisani, Shihab Shamma, and Giovanni M Di Liberto. Neural signatures of musical and linguistic interactions during natural song listening. *Hal preprint*, 2024.

[42] van A Van der Schaaf and JH van van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision research*, 36(17):2759–2770, 1996.

[43] Bidisha Sharma, Xiaoxue Gao, Karthika Vijayan, Xiaohai Tian, and Haizhou Li. Nhss: A speech and singing parallel database. *Speech Communication*, 133:9–22, 2021.

[44] Michael J Crosse, Giovanni M Di Liberto, Adam Bednar, and Edmund C Lalor. The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, 10:604, 2016.

A Perceptual alignment in image-pixel domain

Opposite to latent diffusion, for diffusion models operating on image pixels or mel-spectrograms encodings of natural images and sound, it has been shown in [14, 15] that a hierarchical alignment between coarse/fine structures and low/high frequencies is enforced by the power-law distribution of such data [42]. This law states that spectral power densities decrease as a power of the frequency. Combined with the SNR-properties of diffusion processes, [14] therefore argues that operating in the natural data case, diffusion process generation (or IC estimation) can be viewed as an autoregression in the frequency domain. Furthermore, diffusion models[17, 18] typically produce high-fidelity denoising for intermediate noise levels, due to the inverted U-shaped properties of the loss and modern noise-schedules[17, 18], and, therefore, effectively produce high-fidelity results for frequencies that are not too high. Since human perception is more sensitive to low-frequency content than high-frequency content, [14] hypothesize that the autoregressive inductive bias plays an important role in the success of diffusion models. [15] finds that using a noise process that removes information from all frequencies uniformly can perform equally well; however, a noise process that removes information from low frequencies and then high frequencies performs substantially worse. This shows that the order in which data features appear in the noise process plays an important role in the (autoregressive) generation process of diffusion models. However, it remains less explored how these insights transfer to latent diffusion and if imposing a certain perceptual hierarchy improves such models' performance on tasks like surprisal estimation.

B Model selection

In the following, we detail the model selection procedure for the autoencoder and the autoregressive latent rectified flow model. As we care for both showing hierarchical alignment between perceptually important features and coarse structures, and ultimately the downstream task of estimating surprisal in music, we train autoencoders with different settings and inspect their reconstruction qualities and pitch surprisal estimation capabilities.

B.1 Autoencoder

A practical challenge when training latent representations z with added noise is that the variance of z can grow to counteract the noise and encode all information in coarse structures. For variational autoencoders, such as the one used in [16], the Kullback-Leibler loss term hinders the latents from deviating from the standard normal distribution. For the CAE, the latents cannot grow unbounded due to the use of a hyperbolic tangent (TanH) bottleneck activation that keeps the latents within a range of $[-1, 1]$. Nevertheless, we observe that in the experiments of [16] and using a TanH bottleneck activation for the CAE, the overall variances of representations grow in scales of tenths, thereby silently increasing the SNR during training. This effectively lowers the effect of the noise. In addition to using the original hyperbolic tangent bottleneck of the CAE, we also replace it with a layer norm (LayerNorm) [29], which fixes the variance to that of the noise distribution. In that case, the expected SNR remains constant during training, and the latent noise process is identical to the noise process used for rectified flow latent diffusion except for the noise schedule.

We finetune the publicly available Music2Latent checkpoint of [19] using the same data, architecture, and hyperparameters as described in the paper, except for following a constant, consistency-step schedule with step size initialized to the final value of the pre-trained model. We fix the logit-normal's scaling parameter to $s = 1$ and vary the m parameter (higher m implies more noise is added). For TanH, we use $m = -1, 0$. For LayerNorm, we use $m = -2, -1$. For all models, we then run the same experiments as described in Section 4 and report the results in Figure 3 as E, D. Additionally, we report the results for unaligned autoencoders in the two scenarios: 1) where the training-inference discrepancy is fixed by freezing the encoder and training the decoder with noised latents, reported as $NT = D$, and 2) without the correction, reported as $NT = \emptyset$. For the former, we try several different noise schedules using $m = -2, -1, 0$, since we are interested in ablating against the best possible noise adaptation.

Comparing LayerNorm for aligned representations $NT = D, E$ with unaligned $NT = D$, we find that the perceptual metrics are better for aligned models or similar, most drastically on SI-SDR.

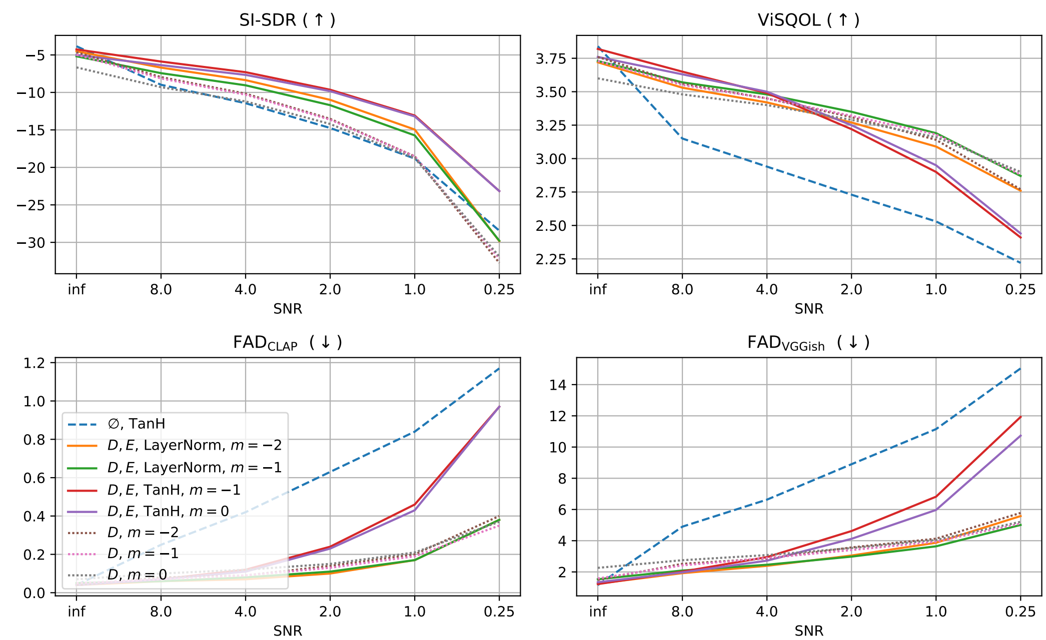

Figure 3: SI-SDR, ViSQOL, $\mathrm{FAD}_{\mathrm{CLAP}}$ and $\mathrm{FAD}_{\mathrm{VGGish}}$, where encoder and decoder are trained with noised-latents (D, E), only decoder (D), and the base model $(\emptyset)$. We show this using the original encoder bottleneck activation of the CAE (TanH) and an alternative (LayerNorm), with fixed latent variance. We provide results for two different noise levels, specified by the logit-normal's mean value m (where lower values correspond to more noise).

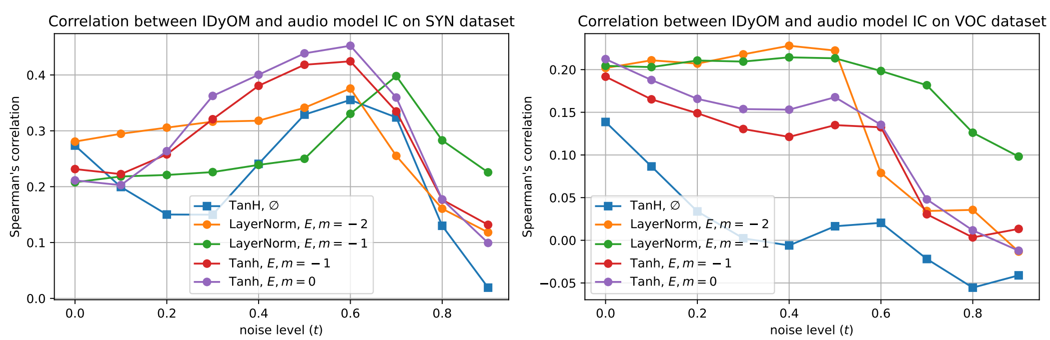

Figure 4: Correlation with IDyOM pitch surprisal for models trained with different latent noise strengths and different bottleneck activation functions.

Comparing the TanH variants against LayerNorm, it is observed that LayerNorm performs better at low SNR, except for on the SI-SDR metric. The FAD metrics for TanH reveal that, at these low SNRs, the reconstructions are unrealistic, even more than the noise-adapted ones.

B.2 Autoregressive diffusion model training details

For the four different autoencoders, we additionally train autoregressive models for our musical surprisal estimation task.

For training the autoregressive rectified flow models in the different latent spaces, we scale the latents to have the same overall variance and use the same data, model, and hyperparameters as in [13] except for lowering the maximum sequence length to 3125, corresponding to ∼5 minutes of audio, running an AdamW optimizer with base learning rate of 10^{-4} with cosine learning rate schedule of 750k steps with a linear warmup of 10k steps, and applying a logit-normal schedule with scaling

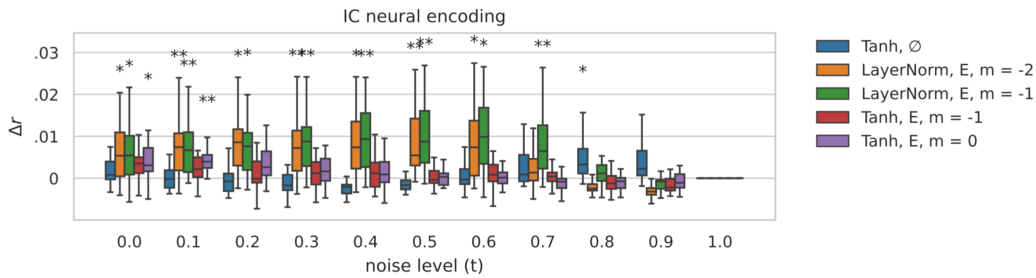

Figure 5: Neural encoding of ICs computed for different models and noise levels. Δr denotes the increase in prediction accuracy when comparing a full model (IC + acoustic envelope) with a reduced model including only the envelope. Bar plots report the mean ± SE across participants (median across electrodes, average across trials).

parameters $m, s = 0, 1$. For our experiments involving musical surprisal estimation in singing voices, we finetune our model on a small private dataset of singing voices running for 36k steps, with a base learning rate of 5×10^{-5} and a warmup of 12k steps.

We then conduct the experiments of Section 5 for all models, and report the results in Figure 4. For the aligned LayerNorm autoencoder models in particular, but also for the TanH models on VOC, we observe a concave downwards shape, where the correlation increases with increasing noise levels until some moderate noise level, and then decreases for extreme noise levels. This indicates that most information relevant to pitch suprisal estimation is present in course structures at these noise levels. This is not the case for the model operating with unaligned representations. Comparing LayerNorm and TanH, it is seen that the highest correlations for the synthetic data (SYN) are found for TanH, whereas for the singing voices data (VOC), they are found for LayerNorm. Interestingly, for LayerNorm, we find that lower SNR-noise schedules push the knee of the curves towards lower noise levels, suggesting that the pitch information is contained in even coarser structures. This is not the case for TanH.

B.3 Neural encoding analysis

The data used in this study are the same as in our previous work [40, 41], where 64-channel EEG responses were recorded from twenty adult individuals as they listened to 18 English songs extracted from the *NHSS Speech and Singing Parallel Database* [43]. For more details about the stimuli, experimental design, data acquisition, and preprocessing, please refer to [41].

In the present study, we aimed to quantify how much variance in brain responses can be explained by musical surprisal as modeled by information content, *i.e.,* its neural encoding. To this end, we used Ridge regression to model EEG responses as a linear combination of two predictors: (i) IC and (ii) the acoustic envelope of the waveform, computed as the absolute value of its Hilbert transform. The envelope served as a nuisance regressor, absorbing variance due to low-level acoustic features and trivial voiced/unvoiced responses, thereby enabling us to isolate the encoding of the higher-level processes related to expectations. To further control for acoustic confounds, unvoiced IC segments were interpolated with constant values sampled from a distribution of ICs estimated for each song.

For each participant and condition, the channel-specific mappings between predictors and EEG were estimated by solving a regularized linear regression problem [44]. Thus, separate and independent optimal filters were estimated for each of the 64 channels, 20 participants, 5 models, and 10 noise levels. Non-instantaneous interactions were captured by including multiple stimulus–response time lags within a $[-100, 700]$ ms window, with an additional 50 ms margin to avoid edge artifacts. Model performance was evaluated using leave-one-out cross-validation across trials, and quantified as the Pearson correlation (r) between the predicted and observed EEG signals at each electrode. The significance of the IC contribution was assessed by comparing the predictive power of a full model (IC + envelope) with that of a reduced model (envelope only). The difference in predictive power (Δr) provides a measure of unique variance explained by IC beyond low-level acoustics.

Statistical analyses were performed using two-tailed t-tests or non-parametric Wilcoxon signed-rank tests for pair-wise comparisons. The choice of using either the parametric or non-parametric test was based on the normality of the data, which was assessed via the Anderson-Darling test. Correction for multiple comparisons was applied where necessary via the false discovery rate (FDR) approach for topographies, and via the Bonferroni correction otherwise.

We report the gains for the four different model variants as a barplot in Figure 5, where stars over the bars indicate significance. Generally, it is seen that both LayerNorm variants have higher gains that are significantly more often than the TanH variants at moderate noise levels above 0.8. These results are consistent with those above for pitch surprisal estimation.

B.4 Conclusion

Due to the mostly superior performance of the most heavily noised LayerNorm variant ($m = -1$) across our experiments, we select it when reporting results in the main manuscript. However, we note that the other variants often outperform the unaligned variant.