Epistemic Inadequacies in Neural Audio Synthesis Evaluation

Riccardo Ancona *
Department of Arts
University of Bologna
Italy
riccardo.ancona@unibo.com

Abstract

Neural audio synthesis technologies have rapidly become ubiquitous in the music and multimedia industries, causing widespread sociocultural, economic, and aesthetic effects. The evaluation stage is a particularly relevant phase of their development, as it defines the implicit aesthetic goals a model should pursue. This study examines sixty-eight recently designed neural audio synthesis models, critically inquiring into their objective and subjective evaluation criteria. Reductionist assumptions, incomplete information, limited experimental replicability, and small sample sizes significantly compromise the epistemic reliability of current evaluation practices. Similarity to pre-existing music and vague notions of quality and musicality emerge as underlying, poorly articulated aesthetic principles. Neural audio engineering inherits its evaluation methodologies from pre-existing, non-creative research fields, uncritically applying them to aesthetic tasks such as automatic music generation. Consequently, the role of listening is merely reduced to numerical grading, rejecting any proper qualitative assessment and flattening the aesthetic experience into uni-dimensional quantitative metrics. In order to overcome these epistemic inadequacies, engineers and musicologists should cooperate to complement thorough objective measurements with phenomenologically-grounded forms of listening.

1 CONTEXT

Neural audio synthesis is a set of deep learning techniques aimed at the automatic generation of sound and music. Most of current AI music is produced employing some form of neural audio synthesis, which has largely replaced previous methods based on the generation of symbolic information. It is an inherently referential and intermedial technique: during the training of a neural audio synthesis model, large datasets of pre-existing music – typically unlicensed – are required, while the generation is conditioned by user inputs in the form of sounds, texts, images, or videos. The ease with which new music can be mass-produced, even by unskilled users, makes it a highly attractive technique for the contemporary media industry, which is investing billions in its further development. Neural audio synthesis is rapidly entering in the workflow of musicians and audio engineers, while also replacing productive segments in the sector, such as in advertising and library music. In reshaping what music-making and music listening are, AI music based on neural audio synthesis is stimulating the emergence of new musical cultures, as well as attracting strong criticisms about its extractivist nature.[1] There is an urgent need to develop critical frameworks to assess the ethic and aesthetic implications in the adoption of this practice (Sturm et al., 2024). The entire production chain must be analyzed empirically, from data collection to model development and distribution, complementing the engineers' work with sociological and musicological insights. Whether in academia or industry, research groups developing neural audio synthesis technologies tend to consist solely of engineers and data scientists, rarely possessing an academic background in music studies (Ancona, 2025). Their vision of what music is and how it should be conceived, created, and evaluated inevitably has a significant influence on the affordance of the technical objects they devise. Following a Science and Technology Studies approach, it is important to trace back the epistemic criteria with which neural audio synthesis models are produced and tested, observing underlying norms and assumptions. Neural audio synthesis as a technique emerges at the intersection of two larger research fields, Music Information Retrieval and Deep Learning, from which it inherits methodologies and worldviews. Both fields aim at developing models and techniques whose efficacy has to be demonstrated objectively. Benchmarks and metrics are thus constantly devised as a way to evaluate the capability of old and new models to compete in specific tasks. There is, as often happens in engineering, a socially-constructed negotiation on which metrics are supposed to better indicate the quality of the models. Different benchmarks contend the status of objectivity and measurement appropriateness according to differing epistemic implications and empirical tests. Metrics themselves are therefore technical objects that have to be critically analyzed to understand how their usage might affect musical cultures. At the current state, the collective discussion surrounding the evaluation methods of neural audio synthesis is merely technical, lacking aesthetic considerations on how the adoption of certain methodologies might affect the sensible results of the models. This article is thus an attempt at critically examining the current state of neural audio synthesis evaluation, finding epistemological inconsistencies and methodological issues, with the intention of stimulating a larger debate on the matter through a humanistic lens.

2 METHODOLOGY

A corpus of sixty-eight neural audio synthesis models released between 2024 and 2025 has been examined, tracking down which evaluation metrics have been employed

*riccardoancona.com

[1] See, for instance, how (Kanhov et al., 2024) analyzed the impact of AI music on traditional Irish music.

in their development. Technical considerations and design choices stated by the engineers authoring these models have been annotated, in order to understand the motivations behind the adoption of evaluation criteria. The corpus includes both academic and industrial research, whose results have been shared in international conferences. Fully closed models to which there are no available technical documents describing their design could not be taken into account for lack of information. While the corpus does not exhaust all the models that have been released or tested in the two years analyzed, it is broad enough to provide a picture of the current state of neural audio synthesis research. Whereas countless research papers are shared everyday on platforms such as arXiv without peer review to show the advancements in the field, this study is exclusively focused on articles accepted on renowned conferences dealing with information engineering, deep learning, and music information retrieval. The acceptance of the research papers in such conferences is a sign that the methodological criteria employed for the evaluation process are considered appropriate to the neural audio research community. The corpus is therefore composed of sixty-eight research articles on neural audio synthesis presented at the International Society for Music Information Retrieval (ISMIR, 2024 and 2025), the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP, 2024 and 2025), the ACM International Conference on Multimedia (ACM MM, 2024 and 2025), and the International Conference on Artificial Intelligence and Machine Learning for Audio of the Audio Engineering Society (AES AIMLA, 2025). The complete list of sources is reported in section A. Given the dynamism of this field, methodologies can radically change in a short time span; thus the choice of studying only the two most recent available years. A comprehensive variety of neural audio tasks can be found in the corpus: some of the most occurring are text-to-audio or text-to-music synthesis (present in 25 papers), video-to-music (12), audio-to-audio generation and accompaniment (10), symbolic data to music (6), instrumental resynthesis (5), timbral transfer (3), foley generation (3), and vocal synthesis (2). Some articles focused on specific architectural issues related to the application of neural networks to audio tasks (8 papers), while others aimed at improving pre-existing techniques in the direction of spatial audio (4), long-form generation (3), and multi-stem control (3). Neural speech synthesis has not been taken into exam in this research, as it typically addresses communication tasks rather than creative sonic processes. Regardless of the wide diversity of applications, most of the papers in the corpus adhere to similar evaluation standards and expose them in similar forms: articles on audio engineering typically follow a very rigid formal structure, which requires a section titled "evaluation" at the end of the paper, right before optional ablation studies and the conclusion. This section is often remarkably shorter than the ones that precede it and is meant to corroborate the validity of the model by comparing it with pre-existing ones (the so-called baseline), usually through the use of tables to show promising numerical results. The majority of the articles contain somewhere in the abstract or in the introduction a sentence stating that the proposed model outperforms the baseline in both objective and subjective metrics – it appears as if every single model beats its contenders, in a paradoxical competition in which every model is entitled to be defined the best in its league. The dichotomy between subjectivity and objectivity is also interesting to note, as it seems to be a praxis of neural audio research to divide the evaluation procedure in these two categories. Objective metrics involve the use of algorithms to automatically compare the output of a model with other models or with studio recorded music; the status of objectivity is thus seen to derive from the computational nature of the process. Subjective metrics are based on listening tests conducted with a group of participants, taken as a representative sample of human listeners. They are typically limited to quantitative measurements, obtained by the average scores given by the participants to sound examples according to predefined aesthetic criteria. They are also inherently comparative, as they only track how much a set of generated sounds is supposed to be better than others, but not what those sounds are in themselves. Qualitative analysis is almost entirely absent, banished from the field as it does not grant numerical objectivity. Given the methodological differentiation between quantitative and qualitative metrics, they have to be addressed separately before drawing general conclusions.

Table 1: Most frequent evaluation metrics in the corpus

Metric	*n*
Fréchet Audio Distance (FAD)	51
Kullback-Leibler Divergence (KLD)	28
Likert scale on aesthetic properties (e.g. 'quality')	25
Contrastive Language-Audio Pretraining (CLAP)	24
Mean Opinion Score (MOS)	17
Fréchet Distance (FD)	15
Inception Score (IS)	14
Mel-cepstral distortion (MCD)	4
Visual assessment of spectrograms	3
Mean-Squared Error	3
Fréchet Inception Distance	3

3 OBJECTIVE METRICS

In the context of neural audio synthesis research, objective metrics are seen as a form of computational objectivity: if a given measurement on a set of sounds results in a value lower or higher than the same measurement performed on a different set of sounds, then there is an established relation granting a mean of unbiased comparison. Deep learning benchmarks usually assume a ground truth, an objective observation of external reality to which the inferred data can be reasonably compared. However, generative AI poses a significant challenge to this paradigm: what is the ground truth when the task is creative generation? In order to compare the product of their research with a ground truth, engineers have to make a strong assumption on what can be considered a good example of music creativity. Neural audio synthesis researchers solve this issue by selecting a pool of studio recorded pre-existing music as a comparative example to what the models are supposed to generate. Any assertion of quality is thus based on the similarity between the model's results and the pre-existing music. To find a ground truth, neural synthesis evaluation sacrifices the possibility of conceiving musical creation beyond an imitative paradigm. If the efficacy of a model is based on its resemblance to an idealized prior example, the way such resemblance is measured becomes extremely relevant. The corpus analyzed in this article shows that the establishment of metrics is the product of a constant negotiation. At least sixty-five different metrics are employed, six of which have been proposed by the authors themselves as a way to reconfigure what the evaluation process entails.

An analysis of their frequency of employment across the corpus shows that there are a few authoritative measures and a long tail of less used techniques, some of which are specific to the task they are assessing. The most used metrics are the Frechét Audio Distance (FAD, computed in 51 papers), the Kullback-Leibler divergence (KL, 28), the Contrastive Language-Audio Pre-training score (CLAP, 24), the Frechét Distance (FD, 15), and the Inception Score (IS, 14). They are all comparative metrics. FAD is by far the most frequent one – while this could point to a collective adherence to a well-defined and shared system of measurement, looking at how it is computed in the corpus rather suggests epistemic arbitrariness.

3.1 Fréchet Audio Distance and Kullback-Leibler Divergence

Originally designed for the evaluation of music enhancement algorithms (i. e. denoising), FAD has been inherited by contemporary neural audio engineering, which keeps using it for a substantially different task. Denoising can be easily reduced to a matter of fidelity to a given standard, while sound generation is a much wider and complex matter. The appeal of FAD in neural audio derives from its capability of comparing similarity without requiring a specific ground truth reference, as it does not work at the level of individual audio tracks. Instead, FAD compares how different a set of AI-generated audio files is from another set of possibly unrelated, non-AI-generated recordings, assuming that each set can be understood as a multivariate Gaussian distribution. The authors of the original FAD algorithm define it a "reference-free free metric" (Kilgour et al., 2019), even though the measurement is always referential to another set of pre-existing audio files. FAD does not directly analyze audio. It uses an embedding algorithm that takes the audio sets as input and converts it into a set of compact numerical vectors (embeddings). In the original implementation, such embedding algorithm, called VGGish, is the encoder of a neural network classifier.[2] Once the audio sets have been converted into two groups of vectors, FAD measures the Fréchet distance on their distribution, which is meant to compute the sonic similarity between the two sets. Several authors in the corpus assume that FAD is an objective measure of perceptual quality.[3] To put it in the words of Maman et al. (2024), "the assumption is that perceptually similar audio snippets yield closely spaced embedding vectors". While a small distance between two sets might effectively mean that their signals are somewhat similar, this does not grant that such similarity is equated on a perceptive level. The similarity of two signals does not necessarily correspond to a sonorous experience of resemblance. A phenomenological experience of sounds could quickly reveal that similar waveforms often sound unrelated, while widely different signals might sometimes be perceived as close, as they are framed similarly for semantic and contextual reasons; yet, phenomenological analysis is not an accepted method of evaluation in neural audio synthesis, as it does not grant objectivity. Given that perceptual similarity is intrinsically subjective, as sounds are organized by individuals according to acquired ontogenetic schemata [4], the assumption of objectivity of FAD is epistemically unsubstantiated. Subsuming qualitative perception to a numerical metric is a reductionist operation. As empirically observed in relation to FAD and other "distribution distance metrics", although with methods that will be criticized in the following section: "no previously used objective metric captures the perceptual quality of the synthesized sounds sufficiently well [...] any quality ranking based on these objective metrics is questionable [...] measuring generator performance is insufficient to measure audio quality"(Vinay and Lerch, 2022). The similarity between embeddings is only indirectly related to a similarity between the audio signals: embedding algorithms are themselves neural networks trained on specific datasets, which they learn to organize into a multidimensional latent space in a way that is both ungraspable and contingent to the training setting. The VGGish embedding algorithm originates from a past generation of machine learning research, in which neural audio synthesis was just at its beginning; it has therefore never observed AI-generated music in its dataset, making it an improper choice as a classifier. It has no notion of the type of signals it is fed to, and for this reason it might reduce them to pre-existing categories without great discerning capabilities. Gui et al. (2024) highlighted the inadequacy of VGGish when in their tests they observed that "four out of five samples with the highest FAD scores consist of bagpipe music", meaning that "particularly recognizable classes may produce unique embedding activations" resulting in higher FAD scores. For this reason, many of the articles in the analyzed corpus employ a different embedding algorithm, such as PANN, Trill, and CLAP.[5] However, the majority of models still explicitly rely on VGGish, while twenty-four articles do not specify which algorithms have been chosen, possibly pointing to the implicit use of VGGish. Gui et al. (2024) also noted that regardless of the choice of embedding algorithm, other parameters related to FAD measurement have a significant impact on the results. Increasing the number of samples included in the set tends to produce progressively better FAD scores. Many articles in the corpus are not very specific in explaining which parameters have been chosen and for which reason. The extreme variability of sample sizes, types of embeddings, and reference audio sets renders FAD scores mostly incomparable between different measurements. Furthermore, most of the models are not open source, so that other researchers cannot perform additional empirical evaluations. This completely collapses the fundamental notion of experimental replicability that lies at the core of scientific research. The most widespread and widely accepted objective metric in neural audio synthesis evaluation is thus neither objective nor reliable. The Kullback-Leibler divergence is not much different in this sense: like FAD, it is a distribution distance metric based on embedded data. Varying the embedding algorithm and the other parameters it is possible to obtain more or less favorable results. It still assumes that statistical similarity equates perceptual quality. If a malicious researcher decided to provide arbitrary values instead of FAD and KL measurements to make it look like their closed source model is superior to its competitors, there would be no easy

[2]The original TensorFlow implementation of VGGish has been in released in conjunction with Audio Set, a dataset for object detection (Gemmeke et al., 2017).

[3]See, for instance, (Mancusi et al., 2025).

[4]The concept of schema has largely been explored on Auditory Scene Analysis. It is based on neurobiological studies on the subjective, ontogenetic acquisition of top-down cognitive structures for source identification. See, for instance, (Yost, 2008).

[5]See (Tailleur et al., 2024) for experimental data on how different embeddings affect FAD. Note that their assessment is based on subjective evaluation methodologies criticized in the following section. FAD has been further critiqued by (Huang et al., 2025).

way to discover that. Metrics that can be so easily manipulated do not convey objectivity and epistemic validity.

3.2 Contrastive Language-Audio Pre-Training Score

CLAP is a pre-trained multimodal algorithm that is able to encode both text and sounds in a shared latent space (Wu et al., 2023). The CLAP score of a sound with respect to its textual prompt is the cosine similarity between the two encoded vectors: the closer the two vectors are in the latent space, the most semantically connected they are supposed to be. Neural audio engineers use CLAP score as an objective measure to assess the semantic adherence between given textual prompts and the resulting sounds in text-to-audio models. They assume that CLAP is inherently capable of capturing connotative nuances relating sounds with their verbal descriptions. However, looking at how CLAP is trained, some doubts might emerge regarding this assumption. Most of the data employed for its training comes from LAION-Audio-630K, an open dataset that collects online available sounds from multiple sources. By far the most used source in the dataset is Freesound, an online platform for sound sharing. Sounds scraped from Freesound have textual descriptions given by the users who uploaded them. An analysis suggests that the dataset might be polluted by all sorts of improper captions, such as promotional links and biographic data of the users (Ancona, 2024). Other sources of training data employed for CLAP include keyword-based datasets that have been artificially augmented to captions by language models. Given the large amount of necessary data to train an algorithm such as CLAP, it cannot rely on properly captioned data produced by professional annotators; instead, it has to establish indirect, automated relationships, whose effective semantic adherence might vary significantly across the dataset. Using CLAP score as an objective metric can thus be a generic way to assess the statistical similarity between textual captions and sounds, but it does not grant an epistemically grounded measure of semantic adherence capable of representing connotative aspects of multimodal relations.

4 SUBJECTIVE METRICS

While neural audio engineers seem to appreciate objective metrics despite their epistemic limitations, they tend to be much more skeptical with regards to subjective analyses. In a recent paper, Meta researchers define human listeners as "inconsistent and resource intensive", proposing instead an automatic system of aesthetic assessment (Tjandra et al., 2025). The incommensurability of subjective judgment is perceived as an insurmountable limit to a metric-based culture. If it is not computable, then it is not reliable. Yet, aesthetic experiences are not so much a matter of numbers than rather an entangled conundrum of neurobiology, contingency, personal memory, and sociocultural context. Engineers try to solve this tension by flattening down the complexity of aesthetics and perception into statistical quantities. The only type of subjectivity allowed in this field of research is an aggregated average: a vague plurality, rather than a real subjectivity. In order to obtain an average of the aesthetic experience, neural audio researchers cannot rely on qualitative, descriptive assessments from listeners, opting instead to quantifiable metrics. Twenty-two out of sixty-eight articles in the corpus do not attempt to perform subjective evaluations, trusting only objective metrics to establish the efficacy of their models. The remaining forty-six show a rather homogeneous context: with the exception of three papers analyzing spectrograms, two conducting ABX trials, and a paper listening to potential memorizations, the remaining articles all employ the same family of subjective evaluation methodologies. Likert scales and Mean Opinion Scores are ubiquitous in neural audio synthesis. The Likert scale is a classic psychometric measurement and possibly the simplest form of quantification of perceived features. It asks participants to rate a phenomenon in an evenly spaced numerical scale, whose range can vary – the most frequent chosen scales in the corpus being 5-point, 10-point, and 100-point. Given that the tests typically aim at capturing multiple different parameters of the listening experience, multiple Likert scales are used in parallel. The Mean Opinion Score (MOS) is the arithmetic mean between multiple measurements, representing them with a single scalar value. Neural audio research inherits this type of testing from telecommunication engineering: several articles in the corpus openly cite the ITU-R BS.1534 recommendation, also called MUSHRA, as a reference (Union, 2015). MUSHRA is an assessment protocol of the International Telecommunication Union (ITU) designed to evaluate middle-quality encoding algorithms and it is therefore meant to help to discriminate the effects of compression to the fidelity of audio transmission. Even though deep learning encoder-decoder architectures have much in common with compression algorithms, what neural audio engineers try to measure with MOS is profoundly different to what ITU defined as the appropriate application of MUSHRA. While telecommunication engineering is exclusively interested in the fidelity of transmission, generative tasks require to judge the quality and musicality of sounds. The listening setting defined by MUSHRA is designed for a specific necessity, which should not be translated to neural audio synthesis without reconsidering some essential issues. For instance, MUSHRA states that "where the conditions of listening tests are tightly controlled [...] no more than 20 assessors are often sufficient for drawing appropriate conclusions from the test" (Union, 2015). While this might be reasonable for audio fidelity, whose only interest is understanding to which extent an average listener is disturbed by minor distortions in a signal, it is a strikingly small sample size to draw any conclusion on more complex aspects of listening. Yet, the most frequent number of participants to subjective evaluation tests contained in the corpus is exactly twenty. Only six out of forty-six tests enrolled more than thirty participants, and twelve enrolled ten or less. Seven articles do not mention the number of participants, compromising any type of epistemic reliability of their tests. In general, the small number of listeners in subjective evaluation tests seems to be completely inadequate to the type of aesthetic parameters they are required to judge. It is hard, if not impossible, to draw any sort of reasonable generalization about qualitative listening experiences from such small samples. Researchers often justify the restricted number of participants with a lack of economic resources. While this might be true in some occasions, AI research is currently the most funded research field. Maybe it is not so much a matter of overall available resources, but rather of resource allocation: reading their research reports, engineers seem to invest only a small amount of their available funds to subjective evaluation. Another aspect of testing inherited by MUSHRA is the length of the audio files presented to the listeners, which it recommends to be "approximately 10s [...] to avoid fatiguing" (Union, 2015). This temporal range is

appropriate when the task is to discern minor timbral and harmonic aberrations in a signal, but it is too short to assess several aspects of music form and temporal coherence. MUSHRA prescribes that the assessors should "have experience in listening to sound in a critical way", but neural audio research rarely specifies the level of listening experience of the participants. In the entirety of the corpus, only a handful of papers give any sort of information regarding who the assessors are, how they have been selected, and in which context the scoring is performed. This lack of information strongly invalidates any attempt of generalization. It is striking how articles can be so much detailed about neural architecture, training parameters, and ablation, while at the same time being incomplete with regard to subjective evaluation procedures. If not intentionally vague, surely such incompleteness manifests a lack of trust or interest in listening tests: MOS measurements are there because they have to, but they often seem to be perceived as ancillary by the authors. Even if neural audio synthesis engineering adopted more accurate subjective evaluation methodologies, selecting larger numbers of assessors and controlling the listening context more rigorously, a major epistemic problem of subjective evaluation would remain unresolved: the choice of subjective criteria to quantify. When assessors are asked to score audio excerpts, they are given predefined parameters, the most ubiquitous ones being "overall audio quality" (present in 33 articles) and "semantic alignment" (25). What overall audio quality means is largely given for granted by the authors and remains unexplained in the corpus, with a few exceptions. Tian et al. (2025) recursively define it as "the rationality and quality of the audio", while Zhang et al. (2025) state that it "measures the naturalness and listenability of the generated audio". The term "quality" therefore seems to attract multiple conflicting attributes, all of which are even more undefined than quality itself. "Semantic alignment" is a much clearer parameter, as it is simply supposed to measure how much the generated audio adheres to the input, being it text, audio, video, or symbolic data; while it is not certain that it can be effectively represented in a linear scale between zero and five, at least its meaning is univocal, even if it requires hermeneutic skills to be interpreted. Quality, as many other parameters selected for subjective evaluation in neural audio synthesis, is inherently fuzzy. Engineers ask assessors to rate "musicality" (3 occurrences), "expressiveness" (3), "plausibility" (2), "naturalness" (2). Even assuming that such categories could be unambiguously defined, reducing them into a linear scale is epistemically inadequate. In no way quality can be reasonably flattened to a uni-dimensional scalar value[6]; a reductionist, deterministic epistemology would at least recognize that it is a multidimensional concept. A musicologically-informed perspective would reject the very possibility of reducing quality to a number, as it is not univocal in its meaning and it is not an inherently comparative measure. Asking if a given sound is qualitatively better than another is a strange request: listening is not measuring performance or efficacy, listening is engaging in a relationship (Voegelin, 2010). The way quality is framed in neural audio synthesis evaluation reveals a lack of knowledge in aesthetics. If neural audio engineering wishes to validate its empirical results through aesthetic judgments, then a thorough education in philosophy of sound, musical aesthetics, and listening hermeneutics is indispensable in its curricula. Another option would be to include musicologists in the research teams, opening the subjective evaluation process to a wider range of possibilities, some of which would necessarily exceed or refuse the domain of rational quantification. A synergy between musicologists and engineers would also help to reframe the underlying objectives of neural audio synthesis research: what is the aim of these models? Which aesthetic criteria are implied? Engineers always tend to opt to some form of rational neutrality, but no criteria is neutral when it comes to aesthetic judgment and subjectivity. Musicologists would also help to better define listening tests when they are supposed to measure musical features such as "temporal consistency" (12 occurrences), "similarity to reference" (5), "harmonic consistency" (3), "spatial correctness" (3), "vocal clarity" (2), "pitch accuracy" (2), and "dynamics accuracy" (1). These concepts need to be thoroughly described before evaluation, and their assessment requires specific hermeneutic skills.

[6] See Wittgenstein (1966), lectures I-IV.

5 CONCLUSIONS

To obtain a representative overview of how model output is evaluated in the field of neural audio synthesis, this study examined sixty-eight models presented at international conferences between 2024 and 2025. The analysis identified significant epistemic inadequacies in both objective and subjective assessment methodologies. Objective metrics employed in the field are almost exclusively based on distributional and comparative measures, obtained by computing distances between embeddings; they represent a significantly indirect way of assessing the similarity between generated audio and reference sounds, yet they are incorrectly defined as measures of perceptual quality. The Fréchet Audio Distance, by far the most common objective metric in neural audio synthesis evaluation, has been inquired with respect to its technical limitations: outdated embedding algorithms and parametric variability undermine its epistemic reliability, especially in the context of closed source models and incomplete technical descriptions. Measurements of Kullback-Leibler divergence suffer from similar issues, while CLAP scores presume that connotative sonic features can be grasped by algorithms trained on highly polluted data. Objective evaluation thus often lacks empirical replicability and bases itself on questionable epistemic assumptions. Subjective evaluation methodologies, on the other hand, are mostly limited to Likert scale ratings represented as Mean Opinion Scores, which are supposed to linearly represent quality and musicality. A highly reductionist and simplistic understanding of musical aesthetics and listening hermeneutics is displayed in how subjective evaluation is conducted. Multidimensional, qualitative features are systematically flattened into scalar uni-dimensional ratings, whose epistemic adequacy is untenable under a musicologically-informed perspective. Qualitative evaluation also suffers from limited numbers of assessors, short listening formats, and vague reporting of essential contextual information. While some research groups demonstrated to be more aware of the inherent methodological limitations of the current evaluation paradigm, providing detailed descriptions of their measurements and questioning important issues, the overall picture observed in this study appears problematic. Neural audio engineers tend to understand musical quality and meaningfulness more as a mere quantitative optimization problem than as an aesthetic, contextual, situated sociocultural matter. Reductionism dominates the field, subsuming listening to numerical grading. To find epistemically grounded evaluation methodologies, a new paradigm should emerge from

a close cooperation between engineers and musicologists. A phenomenological, situated, critical, and discursive perspective needs to be included in the evaluation procedures to complement and overcome reductionism. Neural audio synthesis has yet to develop its own acoustemology[7], a way of knowing through sound – an endeavor that can only be fulfilled by questioning the status quo and radically rethinking what evaluation means in the field of generative AI and what it could mean for the cultures it is fostering.

AUTHOR DECLARATIONS

No generative AI technologies have been employed for the writing of this article.

REFERENCES

Ancona, R. (2024). Una prospettiva critica sui dataset per la sintesi text-to-audio. In *Projecting Memories. 24th Colloquium on Music Informatics*, pages 107–14. AIMI.

Ancona, R. (2025). Ontologies of sound in neural network engineering. In *Proceedings of the 6th Conference on AI Music Creativity*. AIMC.

Feld, S. (2017). On post-ethnomusicology alternatives: Acoustemology. In Giannattasio, F. and Giuriati, G., editors, *Ethnomusicology or Transcultural Musicology?* Fondazione Giorgio Cini.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Channing Moore, R., Plakal, M., and Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In *2017 International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Gui, A., Gamper, H., Braun, S., and Emmanouilidou, D. (2024). Adapting frechet audio distance for generative music evaluation. In *2024 International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Huang, Y., Novack, Z., Saito, K., Shi, J., Watanabe, S., Mitsufuji, Y., Thickstun, J., and Donahue, C. (2025). Aligning text-to-music evaluation with human preferences. In *Proceedings of the 26th ISMIR Conference, Daejeon, Korea,*. ISMIR.

Kanhov, E., Kaila, A. K., and Sturm, B. L. T. (2024). Innovation, data colonialism and ethics: critical reflections on the impacts of AI on irish traditional music. *Journal of New Music Research*, 53(1-2):47–63.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proc. Interspeech 2019*.

Maman, B., Zeitler, J., Müller, M., and Bermano, A. H. (2024). Performance Conditioning for Diffusion-Based Multi-Instrument Music Synthesis. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5045–5049, Seoul, Korea.

Mancusi, M., Halychanskyi, Y., Cheuk, K. W., Moliner, E., Lai, C.-H., Uhlich, S., Koo, J., Martínez-Ramírez, M. A., Liao, W.-H., Fabbro, G., and Mitsufuji, Y. (2025). Latent Diffusion Bridges for Unsupervised Musical Audio Timbre Transfer. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Sturm, B. L. T., Déguernel, K., Huang, R. S., Kaila, A. K., Jääskeläinen, P., Kanhov, E., Cros Vila, L., Cabrera Dalmazzo, D., Casini, L., Bown, O. R., Collins, N., Drott, E., Sterne, J., Holzapfel, A., and Ben-Tal, O. (2024). Ai music studies: Preparing for the coming flood. In *Proceedings of the 5th AI Music Creativity Conference*. AIMC.

Tailleur, M., Lee, J., Lagrange, M., Choi, K., Heller, L. M., Imoto, K., and Okamoto, Y. (2024). Correlation of fréchet audio distance with human perception of environmental audio is embedding dependent. In *32nd European Signal Processing Conference*. EUSIPCO.

Tian, W., Zhu, X., Liu, H., Zhao, Z., Chen, Z., Ding, C., Di, X., Zheng, J., and Xie, L. (2025). DualDub: Video-to-Soundtrack Generation via Joint Speech and Background Audio Synthesis. In *MM '25: Proceedings of the 33rd ACM International Conference on Multimedia*, Dublin, Ireland.

Tjandra, A., Wu, Y. C., Guo, B., Hoffman, J., Ellis, B., Vyas, A., Shi, B., Chen, S., Le, M., Zacharov, N., Wood, C., Lee, A., and Hsu, W. N. (2025). Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv:2502.05139*.

Union, I. T. (2015). *Recommendation ITU-R BS.1534-3 (10/2015). Method for the subjective assessment of intermediate quality level of audio systems*. ITU.

Vinay, A. and Lerch, A. (2022). Evaluating generative audio systems and their metrics. In *Proceedings of the 23nd International Society for Music Information Retrieval Conference*. ISMIR.

Voegelin, S. (2010). *Listening to Noise and Silence. Towards a Philosophy of Sound Art*. Continuum.

Wittgenstein, L. (1966). *Lectures and Conversations on Aesthetics, Psychology and Religious Belief*. University of California Press.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to- caption augmentation. In *2023 International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Yost, W. A. (2008). Perceiving sound sources. In *Auditory Perception of Sound Sources*. Springer.

Zhang, X., Fan, K., Wang, Y., Liang, Y., Lu, J., Du, Z., Shi, Q., and Qin, P. (2025). TAGMO: Temporal Control Audio Generation for Multiple Visual Objects Without Training. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

[7] See (Feld, 2017).

A APPENDIX

Baker, T. and Nistal, J. (2025). LiLAC: A Lightweight Latent ControlNet for Musical Audio Generation. In *Proceedings of the 26th ISMIR Conference, Daejeon, Korea, September 21-25, 2025*, Daejeon, Korea.

Baoueb, T., Liu, H., Fontaine, M., Le Roux, J., and Richard, G. (2024). SpecDiff-GAN: A Spectrally-Shaped Noise Diffusion GAN for Speech and Music Synthesis. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 986–990, Seoul, Korea.

Chang, E., Lin, P.-J., Li, Y., Srinivasan, S., Le Lan, G., Kant, D., Shi, Y., Iandola, F., and Chandra, V. (2024). In-Context Prompt Editing for Conditional Audio Generation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5320–5324, Seoul, Korea.

Chen, K., Wu, Y., Liu, H., Nezhurina, M., Berg-Kirkpatrick, T., and Dubnov, S. (2024). MusicLDM: Enhancing Novelty in text-to-music Generation Using Beat-Synchronous mixup Strategies. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1206–1210, Seoul, Korea.

Cheng, X., Wang, X., Wu, Y., Wang, Y., and Song, R. (2025). LoVA: Long-form Video-to-Audio Generation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Chung, H. C. (2025). AudioGAN: A Compact and Efficient Framework for Real-Time High-Fidelity Text-to-Audio Generation. In *2025 AES International Conference on Machine Learning and Artificial Intelligence for Audio*, London, UK.

Chung, Y., Lee, J., and Nam, J. (2024). T-Foley: A Controllable Waveform-Domain Diffusion Model for Temporal-Event-Guided Foley Sound Synthesis. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6820–6824, Seoul, Korea.

Colombo, M. F., Ronchini, F., Comanducci, L., and Antonacci, F. (2025). MambaFoley: Foley Sound Generation using Selective State-Space Models. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Comunità, M., Zhong, Z., Takahashi, A., Yang, S., Zhao, M., Saito, K., Ikemiya, Y., Shibuya, T., Takahashi, S., and Mitsufuji, Y. (2024). SpecMaskGIT: Masked Generative Modeling of Audio Spectrogram for Efficient Audio Synthesis and Beyond. In *Proceedings of the 25th ISMIR Conference, San Francisco, USA and Online, Nov 10-14, 2024*, pages 420–428, San Francisco, USA.

Concialdi, G., Koudounas, A., Pastor, E., Di Eugenio, B., and Baralis, E. (2024). Ainur: Harmonizing Speed and Quality in Deep Music Generation Through Lyrics-Audio Embeddings. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1146–1150, Seoul, Korea.

Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J., and Pons, J. (2024). Long-Form Music Generation With Latent Diffusion. In *Proceedings of the 25th ISMIR Conference, San Francisco, USA and Online, Nov 10-14, 2024*, pages 429–437, San Francisco, USA.

Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J., and Pons, J. (2025). Stable Audio Open. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Flores Garcia, H., Nieto, O., Salamon, J., Pardo, B., and Seetharaman, P. (2025). Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Guo, H., Zhao, S. Z., Lian, J., Anumanchipalli, G., and Friedland, G. (2024). Enhancing GAN-based Vocoders With Contrastive Learning Under Data-Limited Condition. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, page 480, Seoul, Korea.

Haseeb, M. T., Hammoudeh, A., and Xia, G. (2024). GPT-4 Driven Cinematic Music Generation Through Text Processing. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6995–6999, Seoul, Korea.

He, C., Chen, W., and Wang, M. (2025). Dual Position Attention Time-Frequency Network for Binaural Audio Synthesis. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Heydari, M., Souden, M., Conejo, B., and Atkins, J. (2025). ImmerseDiffusion: A Generative Spatial Audio Latent Diffusion Model. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Hou, S., Liu, S., Yuan, R., Xue, W., Shan, Y., Zhao, M., and Zhang, C. (2025). Editing Music with Melody and Text: Using ControlNet for Diffusion Transformer. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Huang, Z., Luo, D., Wang, J., Liao, H., Li, Z., and Wu, Z. (2025). Rhythmic Foley: A Framework For Seamless Audio-Visual Alignment In Video-to-Audio Synthesis. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Jiang, Y., Chen, Z., Ju, Z., Li, C., Dou, W., and Zhu, J. (2025). FreeAudio: Training-Free Timing Planning for Controllable Long-Form Text-to-Audio Generation. In *MM '25: Proceedings of the 33rd ACM International Conference on Multimedia*, Dublin, Ireland.

Kamath, P., Gupta, C., and Nanayakkara, S. (2025). MorphFader: Enabling Fine-grained Controllable Morphing with Text-to-Audio Models. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Karchkhadze, T., Izadi, M. R., and Dubnov, S. (2025). Simultaneous Music Separation and Generation Using Multi-Track Latent Diffusion Models. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Kim, D., Dong, H.-W., and Jeong, D. (2025a). ViolinDiff: Enhancing Expressive Violin Synthesis with Pitch Bend Conditioning. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Kim, H., Choi, S., and Nam, J. (2024). Expressive Acoustic Guitar Sound Synthesis with an Instrument-Specific Input Representation and Diffusion Outpainting. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7620–7624, Seoul, Korea.

Kim, H., Novack, Z., Xu, W., McAuley, J., and Dong, H.-W. (2025b). Video-Guided Text-to-Music Generation Using Public Domain Movie Collections. In *Proceedings of the 26th ISMIR Conference, Daejeon, Korea, September 21-25, 2025*, Daejeon, Korea.

Kim, K., Koo, J., Lee, S., Joung, H., and Lee, K. (2025c). TokenSynth: A Token-based Neural Synthesizer for Instrument Cloning and Text-to-Instrument. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Kushwaha, S. S., Ma, J., Thomas, M. R. P., Tian, Y., and Bruni, A. (2025). Diff-SAGe: End-to-End Spatial Audio Generation Using Diffusion Models. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Lan, Y.-H., Hsiao, W.-Y., Cheng, H.-C., and Yang, Y.-H. (2024). MusiConGen: Rhythm and Chord Control for Transformer-Based Text-to-Music Generation. In *Proceedings of the 25th ISMIR Conference, San Francisco, USA and Online, Nov 10-14, 2024*, pages 311–318, San Francisco, USA.

Lanzendörfer, L. A., Lu, T., Perraudin, N., Herremans, D., and Wattenhofer, R. (2025). Coarse-to-Fine Text-to-Music Latent Diffusion. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Le Lan, G., Nagaraja, V., Chang, E., Kant, D., Ni, Z., Shi, Y., Iandola, F., and Chandra, V. (2024). Stack-and-Delay: A New Codebook Pattern for Music Generation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 796–800, Seoul, Korea.

Li, X., Zhuo, F., Luo, D., Chen, J., Kang, S., Wu, Z., Jiang, T., Li, Y., Fang, H., and Zhou, Y. (2024). Generating Stereophonic Music with Single-Stage Language Models. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1471–1475, Seoul, Korea.

Majumder, N., Hung, C., Ghosal, D., Hsu, W., Mihalcea, R. F., and Poria, S. (2024). Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization. In *MM '24: Proceedings of the 32nd ACM International Conference on Multimedia*, pages 564–572, Melbourne, Australia.

Maman, B., Zeitler, J., Müller, M., and Bermano, A. H. (2024). Performance Conditioning for Diffusion-Based Multi-Instrument Music Synthesis. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5045–5049, Seoul, Korea.

Mancusi, M., Halychanskyi, Y., Cheuk, K. W., Moliner, E., Lai, C.-H., Uhlich, S., Koo, J., Martínez-Ramírez, M. A., Liao, W.-H., Fabbro, G., and Mitsufuji, Y. (2025). Latent Diffusion Bridges for Unsupervised Musical Audio Timbre Transfer. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Mehta, A., Chauhan, S., and Choudhury, M. (2025). Exploring Adapter Design Tradeoffs for Low Resource Music Generation. In *MM '25: Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10554–10562, Dublin, Ireland.

Michaels, J., Li, J. B., Yao, L., Yu, L., Wood-Doughty, Z., and Metze, F. (2024). Audio-Journey: Open Domain Latent Diffusion Based Text-To-Audio Generation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6960–6965, Seoul, Korea.

Nistal, J., Pasini, M., Aouameur, C., Grachten, M., and Lattner, S. (2024). Diff-a-Riff: Musical Accompaniment Co-Creation via Latent Diffusion Models. In *Proceedings of the 25th ISMIR Conference, San Francisco, USA and Online, Nov 10-14, 2024*, pages 272–280, San Francisco, USA.

Niu, X., Zhang, J., Walder, C., and Martin, C. P. (2024). SoundLoCD: An Efficient Conditional Discrete Contrastive Latent Diffusion Model for Text-to-Sound Generation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 261–265, Seoul, Korea.

Novack, Z., McAuley, J., Berg-Kirkpatrick, T., and Bryan, N. J. (2024). DITTO-2: Distilled Diffusion Inference-Time T-Optimization for Music Generation. In *Proceedings of the 25th ISMIR Conference, San Francisco, USA and Online, Nov 10-14, 2024*, pages 874–881, San Francisco, USA.

O'Reilly, P., Barnett, J., Flores Garcia, H., Chu, A., Pruyne, N., Seetharaman, P., and Pardo, B. (2025). The Rhythm In Anything: Audio-Prompted Drums Generation with Masked Language Modeling. In *Proceedings of the 26th ISMIR Conference, Daejeon, Korea, September 21-25, 2025*, Daejeon, Korea.

Parker, J. D., Spijkervet, J., Kosta, K., Yesiler, F., Kuznetsov, B., Wang, J.-C., Avent, M., Chen, J., and Le, D. (2024). STEMGEN: A Music Generation Model That Listens. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1116–1120, Seoul, Korea.

Pasini, M., Grachten, M., and Lattner, S. (2024). Bass Accompaniment Generation Via Latent Diffusion. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1166–1170, Seoul, Korea.

Postolache, E., Mariani, G., Cosmo, L., Benetos, E., and Rodolà, E. (2024). Generalized Multi-Source Inference for Text Conditioned Music Diffusion Models. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6980–6984, Seoul, Korea.

Qi, A., Xie, X., and Wang, J. (2024). Mtdiffusion: Multi-Task Diffusion Model With Dual-Unet for Foley Sound

Generation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 486–490, Seoul, Korea.

Ren, Y., Li, C., Xu, M., Liang, W., Gu, Y., Chen, R., and Yu, D. (2025). STA-V2A: Video-to-Audio Generation with Semantic and Temporal Alignment. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Rouard, S., Adi, Y., Copet, J., Roebel, A., and Défossez, A. (2024). Audio Conditioning for Music Generation via Discrete Bottleneck Features. In *Proceedings of the 25th ISMIR Conference, San Francisco, USA and Online, Nov 10-14, 2024*, pages 149–153, San Francisco, USA.

Rouard, S., San Roman, R., Adi, Y., and Roebel, A. (2025). MusicGen-Stem: Multi-stem music generation and edition through autoregressive modeling. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Shi, Q., Du, Z., Lu, J., Liang, Y., Zhang, X., Wang, Y., Peng, J., and Yuan, K. (2025). AudioCache: Accelerate Audio Generation With Training-Free Layer Caching. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Streich, G., Lanzendörfer, L. A., Grötschla, F., and Wattenhofer, R. (2025). Generating Vocals from Lyrics and Musical Accompaniment. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Tal, O., Ziv, A., Gat, I., Kreuk, F., and Adi, Y. (2024). Joint Audio and Symbolic Conditioning for Temporally Controlled Text-to-Music Generation. In *Proceedings of the 25th ISMIR Conference, San Francisco, USA and Online, Nov 10-14, 2024*, pages 264–271, San Francisco, USA.

Tang, J., Cooper, E., Wang, X., Yamagishi, J., and Fazekas, G. (2025). Towards An Integrated Approach for Expressive Piano Performance Synthesis from Music Scores. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Tian, W., Zhu, X., Liu, H., Zhao, Z., Chen, Z., Ding, C., Di, X., Zheng, J., and Xie, L. (2025). DualDub: Video-to-Soundtrack Generation via Joint Speech and Background Audio Synthesis. In *MM '25: Proceedings of the 33rd ACM International Conference on Multimedia*, Dublin, Ireland.

Tsai, F. D., Wu, S.-L., Kim, H., Chen, B.-Y., Cheng, H.-C., and Yang, Y.-H. (2024). Audio Prompt Adapter: Unleashing Music Editing Abilities for Text-to-Music With Lightweight Finetuning. In *Proceedings of the 25th ISMIR Conference, San Francisco, USA and Online, Nov 10-14, 2024*, pages 634–641, San Francisco, USA.

Wang, K., Deng, S., Shi, J., Hatzinakos, D., and Tian, Y. (2025a). AV-DiT: Taming Image Diffusion Transformers for Efficient Joint Audio and Video Generation. In *MM '25: Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10486–10495, Dublin, Ireland.

Wang, X., Wang, Y., Wu, Y., Song, R., Tan, X., Chen, Z., Xu, H., and Sui, G. (2024). TiVA: Time-Aligned Video-to-Audio Generation. In *MM '24: Proceedings of the 32nd ACM International Conference on Multimedia*, pages 573–582, Melbourne, Australia.

Wang, Y., Chen, H., Yang, D., Wu, Z., and Wu, X. (2025b). AudioComposer: Towards Fine-grained Audio Generation with Natural Language Descriptions. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Xie, Z., Xu, X., Wu, Z., and Wu, M. (2025). PicoAudio: Enabling Precise Temporal Controllability in Text-to-Audio Generation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Yeh, J.-W., Liu, E. M., and Liu, Y.-W. (2025). Improved Singing Voice Conversion with Frame-Level Content and Melody-Informed Speaker Embeddings Using Cross-Attention. In *2025 AES International Conference on Machine Learning and Artificial Intelligence for Audio*, pages 220–228, London, UK.

You, W., Zuo, H., Wu, J., Zhang, D., Zhou, Z., and Sun, L. (2025a). Spatial-Temporal Decomposition and Alignment in Controllable Video- to-Music Generation. In *MM '25: Proceedings of the 33rd ACM International Conference on Multimedia*, Dublin, Ireland.

You, Y., Wu, X., and Qu, T. (2025b). TA-V2A: Textually Assisted Video-to-Audio Generation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Yuan, Y., Liu, H., Liu, X., Huang, Q., Plumbley, M. D., and Wang, W. (2024). Retrieval-Augmented Text-to-Audio Generation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 581–585, Seoul, Korea.

Zhang, J., Parada, P. P., Jalal, M. A., and Saravanan, K. (2025a). Diffusion based Text-to-Music Generation with Global and Local Text based Conditioning. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Zhang, L. and Fuentes, M. (2025). SONIQUE: Video Background Music Generation Using Unpaired Audio-Visual Data. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Zhang, X., Fan, K., Wang, Y., Liang, Y., Lu, J., Du, Z., Shi, Q., and Qin, P. (2025b). TAGMO: Temporal Control Audio Generation for Multiple Visual Objects Without Training. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Zhang, Y., Xu, X., and Wu, M. (2025c). Smooth-Foley: Creating Continuous Sound for Video-to-Audio Generation Under Semantic Guidance. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.

Zhang, Z. and Akama, T. (2024). Hyperganstrument: Instrument Sound Synthesis and Editing With Pitch-Invariant Hypernetworks. In *2024 IEEE International*

Conference on Acoustics, Speech and Signal Processing, pages 6640–6644, Seoul, Korea.

Zhang, Z., Wu, X., Xu, J., Ma, T., Yao, T., Wu, W., and He, L. (2025d). Temporal-Conditioned Symbolic Alignment for Controllable Text-to- Music Generation. In *MM '25: Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10728–10737, Dublin, Ireland.

Zong, Y. and Reiss, J. (2025). Learning Control of Neural Sound Effects Synthesis from Physically Inspired Models. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India.