Neural Audio Synthesis for Sound Effects: A Scope Review

Mateo Cámara, Fernando Marcos, Anders R. Bargum, Cumhur Erkut,
Joshua Reiss, *Member, IEEE*, José Luis Blanco *Member, IEEE*

***Abstract*—Neural Audio Synthesis is dedicated to generating sound through generative neural networks. Sound effects are defined as auditory elements that complement a specific scene (in cinema, fiction, or videogames), support a storyline, enhance a fictional environment, or improve the perceived plausibility and presence (including Virtual Reality) without being music or dialog. This manuscript presents a quantitative literature review of the literature that intersects these two domains: the neural generation of sound effects. By leveraging large language models, we performed an extensive and systematic survey of the major scientific repositories, filtering the most relevant articles to ensure a thorough analysis. Our study examines various generation paradigms employed in sound synthesis, the specific types of sound effects created, the datasets used, and the evaluation metrics considered. Furthermore, we provide a forward-looking discussion on the evolution of this field towards multimodal approaches, where sound generation might integrate with other sensory modalities. All supporting materials and code are available online.**

***Index Terms*—Neural Audio Synthesis, Sound Effects, Foley Effects, Audio Signal Processing, SFX, Generative synthesis.**

I. Introduction

Neural Audio Synthesis (NAS) refers to generating sound using advanced deep learning algorithms, particularly those involving neural networks [1]. These techniques enable the synthesis of high-quality audio by learning from data; often large corpora of audio samples, but also, in some settings, from very small datasets or even a single example.

NAS has remarkable applications across multiple domains. In music synthesis, it allows generating novel sounds and compositions [2], [3]. In virtual reality, to enhance immersive experiences by providing plausible soundscapes [4]. In interactive media, to produce more engaging and dynamic audio content and complement other modalities [5], [6].

Overall, generative capabilities in NAS represent a significant leap forward in audio synthesis. To avoid ambiguity on neural models structure and training, we distinguish the *generative paradigm* from the *backbone*. The generative paradigm is the learning principle that determines the objective and the sampling rule used to model the data distribution. The backbone is the parametric architecture that instantiates that principle, supplying capacity and inductive bias. Conditioning signals (text, vision, audio, control) are orthogonal and can be attached to either choice [7], as a form to control the generation process.

Mateo Cámara (corresponding author, email: mateo.camara@upm.es), Fernando Marcos, and José Luis Blanco are with the Signal Processing Applications Group and the Information Processing and Telecommunications Center (Universidad Politécnica de Madrid, Spain), Anders R. Bargum and Cuhmur Erkut are with the Multisensory Experience Lab (Aalborg University Copenhagen, Denmark), and Joshua Reiss is with the Centre for Digital Music (Queen Mary University of London, UK).

Compared with pre-NAS methods, NAS learns generative priors from data and supports flexible conditioning, improving scalability, diversity, and controllability. Such capabilities complement rather than replace interpretable Digital Signal Processing (DSP) approaches, including classical methods [8], [9] and more recent Differentiable DSP (DDSP) techniques [3], providing an extended framework for audio practitioners.

The field of NAS is vast. In this study, we narrow our focus specifically to the synthesis of *sound effects*. In doing so, one should disambiguate two closely related but distinct concepts: "audio effects" and "sound effects." Audio effects typically refer to processing techniques applied to existing audio signals to alter their characteristics for aesthetic or stylistic purposes. Common examples include reverberation, vibrato, echo, or distortion, which modify certain acoustic properties while preserving the fundamental nature of the original sound [10]. On the other hand, sound effects denote audio elements designed to enhance plausibility or contribute artistically to a scene, without being musical or spoken in nature [11]. Typical examples are footsteps, door creaks, or ambient noises, usually crafted by Foley artists. Foley, a specialized practice within the audio industry, involves recreating these sound effects by employing everyday objects to match visual content in films, television, and other multimedia productions. This second category—sound effects—is the primary focus of our research.

In this context, NAS has demonstrated significant potential in synthesizing "sound effects." Neural networks, trained on extensive datasets of audio samples, can produce high-quality and highly plausible sounds that align closely with the detailed requirements of Foley synthesis. Unlike traditional methods used in Foley studios, which involve manually recreating sounds to match visuals, NAS allows for the rapid generation of diverse sound variations with minimal effort. This not only enhances plausibility but also drastically reduces the time and labor needed for sound design. Additionally, NAS provides the flexibility to generate sound effects that are difficult or even impossible to create manually, thereby expanding the creative possibilities in audio synthesis.

Despite the promising capabilities of NAS in generating plausible sound effects, one of the main challenges lies in controlling and fine-tuning these systems. While NAS offers significant advantages, its complexity—primarily on the developer side—can be a barrier, requiring specialized knowledge to operate effectively. This complexity may limit its accessibility. "Procedural audio," on the other hand, refers to the real-time generation and manipulation of sound through algorithms and rule-based systems, allowing dynamic adaptation to changing conditions [12]. Although NAS can generate sounds dynamically, its current limitations in intuitive real-time control prevent it from fully aligning with procedural audio principles.

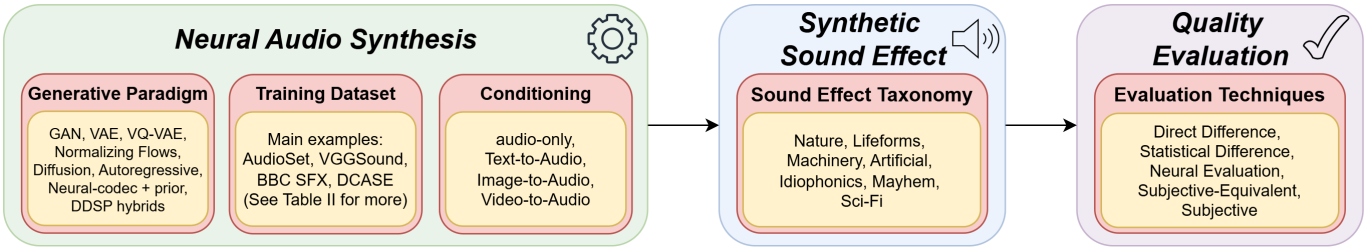

Fig. 1. Conceptual framework of the analysis (from left to right): train for Neural Audio Synthesis (paradigm, datasets, conditioning), synthesize Synthetic Sound Effects (taxonomy), assess output Quality (objective/subjective). Icons mark design, output, and validation.

Thus, achieving real-time adaptability remains a future goal for NAS, representing a benchmark that the technology is gradually approaching but has yet to fulfill. Expanding the real-time capabilities of NAS could pave the way for broader adoption and integration into various creative industries. Closing the control-and-latency gap would turn NAS from an offline black box into an authorable, real-time instrument for interactive pipelines (games, extended reality, or live performance).

In this paper, we provide an overview of the state of the art in the synthesis of sound effects using neural generative models. We discuss the types of generated sound effects, the metrics used to validate them, the datasets used for training, the specific generative paradigms employed, and whether they incorporate any form of multimodality. Figure 1 presents a conceptual framework for the proposed analysis represented as a left-to-right pipeline. First, NAS covers the design choices of the system: generative paradigm, training datasets, and conditioning (image, text, video). These choices produce a Synthetic Sound Effect, organized by a sound effect taxonomy. Finally, Quality Evaluation evaluates results using objective and subjective methods.

To enhance the quality and scope of our overview paper, we designed a pipeline utilizing web scraping and Large Language Models (LLMs) to evaluate as many documents as possible preliminarily. Subsequently, the evaluations and conclusions of the review were conducted personally. Along with this document, we provide the code and the prompts that implement our procedure to ensure reproducibility. These align with the EU guidelines on the responsible use of generative AI in research [13] and preserve the core principles behind the PRISMA framework [14].

The remainder of the paper is as follows: Section II is about the background of the area, Section III exposes the research methodology, Section IV describes the sound taxonomy followed, Section V describes the neural synthesis techniques, Section VI presents the two main tables about the neural sound effects synthesis, Section VII discusses the multimodality in NAS, Section VIII describes the evaluation methods, Section IX discusses the whole quantitate literature review, and Section X concludes the document.

II. Background

Classic audio synthesis is a well-explored topic with numerous significant references that are widely used. Fundamental references are often found in books that serve as cornerstones in the field.

For instance, in *Designing Sound*, [8] Andy Farnell provides a practical guide on constructing sound effects and offers a comprehensive taxonomy of sound synthesis techniques. His book is invaluable for understanding the principles of procedural audio and how to implement various sound effects using computer algorithms. Farnell's work emphasizes the importance of understanding the underlying physics of sound to create more plausible and engaging sound effects.

The Sounding Object, edited by Davide Rocchesso [15], is a significant reference in sound engineering and music synthesis. The book comprehensively explores physical sound synthesis, a key methodology for generating plausible sound effects, and discusses the fundamental principles underlying sound generation from physical models. It includes multiple case studies and practical examples illustrating how material properties and structural forms influence acoustic characteristics. This resource offers valuable theoretical insights and practical guidance relevant to research and application in the field.

Similarly, Ric Viers' *The Sound Effects Bible* [16] is a guide to sound design, detailing the operational mechanisms of various sound effects commonly produced in Foley studios. The book systematically addresses the methodologies for recording and processing these effects, highlighting widely adopted industry techniques and equipment. Viers' work provides practical knowledge essential for sound designers and Foley artists, effectively bridging theoretical concepts and real-world applications.

Perry Cook's *Real Sound Synthesis for Interactive Applications* [9] is another foundational text, particularly relevant for understanding the principles and techniques underlying sound synthesis in interactive environments. Cook's book thoroughly addresses algorithmic approaches for generating sounds that react dynamically to user interactions and real-time inputs. It emphasizes the importance of perceptually informed synthesis methods, offering both theoretical discussions and practical implementations that support interactive audio applications.

These references provide a solid theoretical foundation for working with sound effects in general without delving into the specifics of neural or machine learning-based sound synthesis. They cover a broad range of topics, from the physics of sound and procedural audio to practical Foley techniques, equipping readers with the knowledge needed to create high-quality sound effects. Next, we explore NAS foundamental references:

Moffat and colleagues, in their work *Sound Effect Synthesis*, [17] review techniques for sound effect synthesis used in

creative media, without delving into neural synthesis details. However, they offer a thorough overview of sound effects, aligning with the definition of sound effects we use in our work. Their review is a key reference on the methods and uses of sound-effect synthesis in media.

For the generation of variability and new sound effects, the fundamentals of generative models are central. An accessible review is provided by David Foster's *Generative Deep Learning*, [18], which also includes references on working with sound. This book provides an overview of the principles behind generative models and their applications, including sound synthesis. More specific to the field of NAS, the reference [19] provides a full understanding of how to use generative models in audio.

Combining knowledge from both sound effects and neural synthesis is critical. A relevant work in this area is by Natsiou et al. [20], who review generative models typically used in NAS without focusing exclusively on any specific type of sound. Their work also includes an evaluation of the methods used to assess these systems, providing a comprehensive overview of the current state of generative models in sound synthesis. Another interesting review about deep models for audio synthesis, yet even more recent, can be found in [21].

Additionally, the authors of *The State of the Art in Procedural Audio* [22] present a recent review on procedural synthesis, highlighting relevant aspects of NAS. Although they currently do not classify NAS as procedural audio due to its inherent limitations regarding controllability, their insights underscore the ongoing advancements bringing NAS closer to fulfilling the key criteria of procedural audio. This suggests that NAS is rapidly approaching the necessary conditions for full integration into the procedural audio domain.

The reviews in [23], [24] touch upon NAS-based techniques applied to music and speech, areas that diverge from our primary focus. However, they are still relevant due to the significant overlap between these disciplines and our field of interest. They provide a comprehensive analysis of sound synthesis using neural networks, which is valuable for understanding the broader capabilities and limitations of NAS.

In *Evaluating Generative Audio Systems and Their Metrics*, [25] by Vinay et al., the authors review objective quality metrics for generative sound synthesis. However, these metrics are highly relevant and have been used to categorize the works discussed here. Additionally, the paper mentions subjective metrics evaluated through listening tests, providing a comprehensive framework for assessing generative sound quality.

The author of the doctoral thesis *Deep Learning for the Synthesis of Audio Effects* [26] conducts an in-depth study on how to work with generative models in this context. The thesis focuses on using DDSP and the generation of percussive effects. The author explores how DDSP can create high-fidelity sound effects by leveraging neural networks to model traditional signal processing components. This approach allows for the seamless integration of deep learning techniques with classic audio processing methods, offering a powerful toolkit for sound designers.

Together, these references offer a solid theoretical and practical foundation for understanding both traditional and neural-based sound synthesis techniques. They cover a range of topics from the fundamentals of generative models and procedural audio to the practical applications and evaluation methods of sound effect synthesis, equipping readers with a comprehensive understanding of the field.

While the aforementioned literature provides crucial context—spanning foundational synthesis principles, general generative models, and related audio domains—a notable gap remains. Specifically, a comprehensive review dedicated exclusively to the field of Neural Sound Effect Synthesis is currently lacking. Existing surveys often address broader scopes like general audio synthesis [20], [21], focus on distinct domains such as music or speech [23], [24], cover traditional sound effects techniques devoid of deep learning [17], or examine tangential aspects like procedural audio definitions [22] and evaluation metrics [25]. This fragmentation highlights the necessity for the present work, which aims to consolidate the state-of-the-art, identify key methodologies, challenges, and future directions specifically within sound effects. Such a focused overview is essential for advancing research and application in this rapidly evolving subfield.

III. Research Methodology

A. Definition and Scope

We build on the line opened by WaveNet [27] and SampleRNN [28], which showed that neural models can learn long-range temporal structure directly from raw audio. Magenta's NSynth [2] then helped translate these ideas into creator-facing tools and widely used benchmarks. Together, these works established a practical path for data-driven generation across speech, music, and effects in creative and production settings.

Since then, the application of neural networks in sound synthesis has grown remarkably. As our research focuses specifically on sound effects, (defined in the introduction), this review excludes all sound synthesis strategies outside the categories documented by the BBC[1] [11]. Therefore, music and speech are not considered in our study. Although "music as a sound effect" appears in that taxonomy, we do not include it here because it retains its musical nature.

A notable focus in NAS is on percussive sound synthesis. This stems both from interest in the class and from the fact that phase inaccuracies in percussive sounds are less perceptible to listeners. Hence, evaluations more directly reflect a model's generative ability rather than algorithmic-dependent phase reconstruction. For this reason, we also include systems that synthesize acoustic drum sounds, emphasizing transient energy over pitched content—even though drums are musical instruments.

We restrict our scope to generative methods, excluding traditional non-generative workflows. We emphasize studies that address the core challenges of Foley: the synthesis of complementary sound layers (such as footsteps, cloth, or props) and audio-visual synchronization. Throughout this paper, we refer to this temporal synchronization as *alignment*—specifically,

[1]BBC (1931) distinguishes six primary genres of sound effect: Realistic, confirmatory; Realistic, evocative; Symbolic, evocative; Conventionalised; Impressionistic; Music as an effect.

the frame-accurate matching of audio onsets to visual actions. It is important to note that alignment is independent of the conditioning modality. It can be achieved through both audio-only controls or multimodal (video) conditioning (e.g., [29]).

Conditioning is also a critical factor, with many models steered by image, text, or video. While we acknowledge the rapid growth of multimodality and explicitly catalogue these conditioning signals in our analysis, this review is organized by generative paradigm rather than input modality. Therefore, we discuss audio-only and multimodal methods side-by-side within their corresponding architectural families.

From a modelling standpoint, text-to-audio (TTA) systems form a superset of neural sound-effects synthesis: a TTA model that generates generic environmental audio can produce sound effects as a subset of its outputs. In this review we restrict ourselves to TTA works that explicitly target sound effects, either through their datasets, prompts, or evaluation protocols, and we do not attempt an exhaustive survey of general-purpose TTA models.

B. Identification and Selection of the State of the Art

Because relevant work may appear under heterogeneous labels (e.g., Foley/sound effects, audio effects, VTA, TTA), a narrow query such as *"neural audio synthesis" AND "sound effects"* risks omitting pertinent studies. We therefore deliberately broadened queries and venues to minimize omissions, and used a Large Language Model (LLM) for the first-pass triage (title/abstract understanding) and metadata extraction. Final inclusion and detailed annotations were performed manually, and are described below. This design prioritizes breadth of topic coverage rather than a SOTA ranking while keeping the process reproducible.

This approach allowed a massive survey of the literature, being the most independent to the search keyword, and relying mostly on the understanding of the title and abstract. It is proven that LLMs are proficient at handling non-structured data, as it is represented in a human-readable text [30], that results may vary between LLMs, and that these models indeed incorporate their own limitations [31]. Nevertheless, the authors used it as an assistant for their work to find relevant studies, not as a definitive procedure.

To capture the full range of relevant literature, we conducted a comprehensive search in titles and abstracts using the query:

```
(deep OR neural OR generative model)
AND (audio OR sound OR effects)
AND (synthesis OR generation)
```

This approach aimed to encompass all potential relevant studies. The search terms are interchangeable and can appear together or separately in the title or abstract.

We stored all titles and abstracts, cross-checked them to eliminate duplicates, and applied an initial filter using a LLAMA-3 model[2]. Our screening tools are developed in Python 3.10, and the model runs on a NVIDIA A100 GPU. The model outputs structured fields that are schema-validated and human-verified (with temperature at value zero and fixed prompt version). This model was characterized to determine:

- Whether the topic is related to neural audio synthesis.
- The type of signal it addresses: music, speech, or sound effect.
- The specific generative paradigm employed (details given in Section V): Generative Adversarial Networks (GANs), Variational AutoEncoding (VAE), Vector Quantize-VAE (VQ-VAE)[3], Normalizing Flows, Diffusion, Autoregression, Neural Codec modification, and DDSP.

This filtering allowed us to exclude a substantial number of irrelevant documents based on our criteria. We only kept those that were related to neural audio synthesis and sound effects and employed a neural architecture (a sound effect produced by any other kind of procedure would be removed at this point).

Fig. 2 provides a flowchart of the pipeline that was covered. The actions taken were as follows:

- Use the query in Scopus, Arxiv, and Web of Science in the title, keywords, and abstract. These databases were chosen due to their recognized scientific credibility and extensive indexing of pertinent conference proceedings and peer-reviewed literature, ensuring broad coverage of high-impact research. We did not use Google Scholar search at this stage as it might not be replicable, being a search engine rather than a curated database, which limits its search depth (e.g., to 1000 results). The queries generated lists of 12628, 1601, and 8480 articles, respectively.
- Consolidate the titles, eliminate errors, and remove duplicates, remaining a total of 12562.
- Do a LLAMA-3 screening based on title and abstract. Tagged all entries within three different categories: related to NAS, type of generated signal, type of generation architecture. This process took around 17 hours, about 5 seconds per entry.
- Filter based on first tag, whether it is related to NAS. A total of 6383 documents remained.
- Filter based on second tag, whether it is sound effects related. A total of 137 documents remained. Based on the neural architecture, the third filter was unnecessary as it did not filter out any entry: the first filter was enough to acknowledge if it is neurally powered.
- Check all articles based on a full reading. At this point, we discarded ten articles. A total of 129 were valid at this point.
- Perform forward check (scanning reference lists of included papers) and backward check (using "Cited by" tool for each included paper) in Google Scholar. We manually included ten documents that were missing.
- The final number of scientific articles is 139.

Fig. 3 illustrates the number of papers published on neural audio synthesis up to December 2024. There is a clear upward trend, particularly pronounced in 2023. It is expected to double the number of articles by the end of 2024 (up to 60).

[2]https://ai.meta.com/blog/meta-llama-3/

[3]VQ-VAE is fundamentally different from a standard VAE, being dictionary-based quantization that replaces the VAE posterior and permits sequence priors).

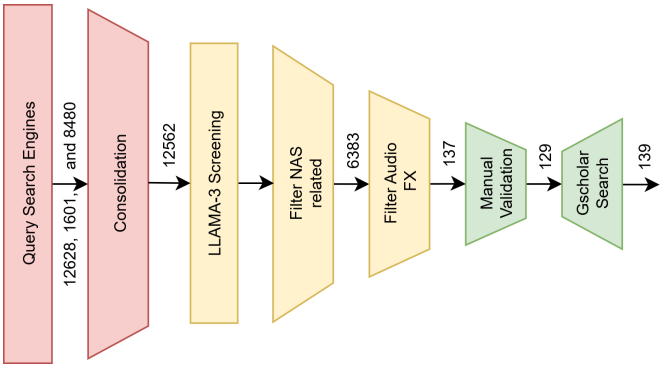

Fig. 2. Flowchart of the actions taken to find relevant scientific publications. In red are steps where an algorithm was used; in yellow, where an AI was used; and in green, where manual steps were taken.

The increase in papers began in 2018, coinciding with the introduction and maturation of enabling technologies such as WaveNet (2016). In [22], it is noted that procedural audio research (not NAS) peaked in 2017 but subsequently declined. Authors argue that this decline could be attributed to the rise of NAS. Our investigation supports this hypothesis, as NAS has experienced significant growth starting from 2018.

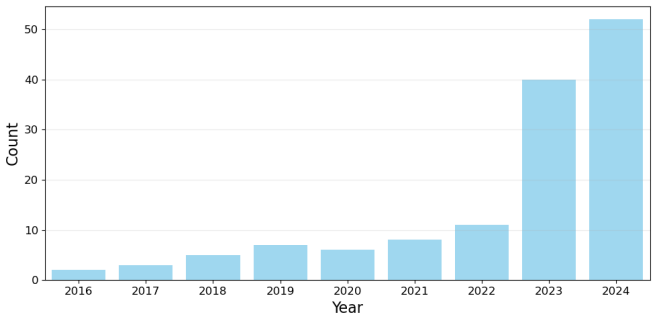

Fig. 3. Number of articles published since 2016 until 2024.

Fig. 4 illustrates the evolution of generative paradigms adoption from 2017 onwards, highlighting a significant trend shift. Initially, the field was dominated by GANs. This trend persisted until around 2022, when diffusion models, initially popularized for image generation, were increasingly adapted for audio synthesis. The growing use of these models marked a new phase in generative techniques. By this time, diffusion systems became the leading tools for creating sound effects, alongside traditional methods like GANs and VAEs.

IV. Sound Taxonomy

The taxonomy used in this study adheres strictly to the one defined in the *Practicals* section of Farnell's book [8]. We are aware that other taxonomies, such as hierarchical ones, might offer advantages in defining interaction types, materials, states of matter, or fundamental details of sound physiology. Despite these potential benefits, we opted for a compact, flat taxonomy similar to that proposed in [22]. Our taxonomy is composed of seven groups:

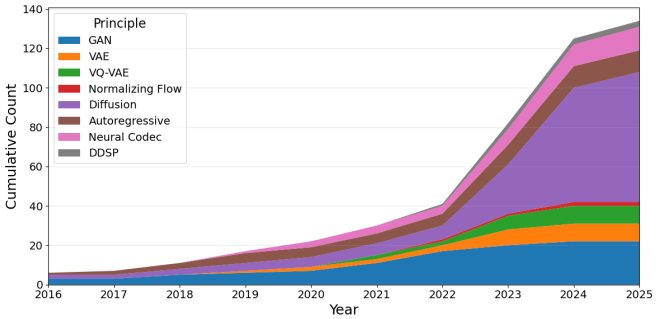

Fig. 4. Neural generative paradigms used to synthesize sound effects per year. A single article may use more than one paradigm.

- **Artificial**: This group consists of human-made elements that are not naturally occurring. It includes various types of electronics, buttons, and alarms. Importantly, machinery is not included here, as it has a dedicated category.
- **Idiophonics**: This category covers the interaction of everyday objects that maintain their structure after impact. This includes unspecified material collisions like crushing, rubbing, scraping, and bouncing. Acoustic drums are also classified here since they do not change shape after being struck.
- **Nature**: This includes sounds that occur naturally, corresponding to the elements of earth, water, air, and fire, as well as the states of matter: solid, liquid, gas, and plasma. Examples are fire crackling, burning wood, flowing water, rain, and thunder.
- **Machinery**: Sounds produced by the operation of machinery, often due to imperfections, fall into this category. This includes the sounds of machines operating, data being transferred through a modem, pneumatics, and relays.
- **Lifeforms**: Any sound generated by the activity of living beings, excluding speech, belongs to this category. This includes footsteps, sounds from terrestrial or marine animals, insects, barking, and snake hissing.
- **Mayhem**: Sounds associated with destruction, aggression, and death are included here. This category covers weapon sounds, explosions, rocketry, and even some sounds that do not exist in reality, such as the exaggerated sound of a gun being pointed in movies.
- **Sci-Fi**: This category encompasses hyper-realistic sounds that extend beyond known real-world phenomena. While some of these sounds might fit into other categories, such as alien machinery, they often lack a real-world reference. Examples include lightsabers, R2D2's sounds, and teleportation effects.

A substantial number of the analyzed studies do not explicitly specify the type of sound effect they generate but do detail the datasets utilized. Therefore, there will be a second categorization based on the datasets used while details on the sound category are not included.

This dual approach allows readers to analyze works based on sound type, which is more practical for application purposes, and based on the dataset, which is more suited for

engineering and technical use. One can assume that if no sound effect is mentioned in their document while a massive dataset is mentioned, all types of recordings were used for training.

V. Neural Generation Paradigms

A. Preliminaries

We denote the observed audio by x (either waveform or a time–frequency representation) and the latent paremeters by z. Most models learn two components: an *inference* map that converts x into a compact representation z (when applicable), and a *generator* that maps z back to audio. Latents can be *continuous* (real vectors) or *discrete* (indices or "tokens"). Discrete latents typically arise from a learned codebook. The resulting token sequence can then be modeled by a separate prior and decoded back to audio. We use *prior* to denote both the latent prior $p(z)$ in continuous models and the sequence prior over discrete tokens in vector-quantized/neural-codec systems. GANs and Diffusion do not learn an explicit $p(z)$ beyond the base noise.

Training objectives across the literature fall into a few simple families. *Likelihood-based* models directly optimize the data likelihood. This is exact in autoregressive models and in normalizing flows, and it is optimized through a variational lower bound in VAEs. *Implicit-likelihood* models, such as GANs, match the data distribution via an adversarial game rather than a tractable likelihood. Diffusion models are trained via denoising or score-matching losses, which connect to likelihood through the score function $\nabla_x \log p_t(x)$.

It is often useful to adopt a *transport* view. Let p_ϵ be a simple base noise distribution and p_{data} the data distribution. Transport-based methods learn a map T_θ (or a velocity field $v_\theta(x,t)$) such that $T_{\theta\#}p_\epsilon \approx p_{\text{data}}$, where $\#$ denotes the push-forward [32]. Normalizing flows realize this transport with invertible maps and provide exact likelihoods. Diffusion models learn a time-dependent transport that inverts a noise process. In what follows, we use *likelihood* to refer to either a tractable $p_\theta(x)$ (exact in autoregressive models and flows) or its variational lower bound (in VAEs). GANs are implicit-likelihood models, and diffusion models are trained with denoising / score-matching objectives. Optimization is stochastic and gradient-based (SGD/Adam) in all cases. Quantizers rely on straight-through estimators. Likelihood models maximize $\sum \log p_\theta(x)$, diffusion models minimize denoising losses, and GANs optimize adversarial objectives.

Conditioning turns an underdetermined generation task into a controllable one: it aligns the output with user intent, context, and timing. A conditioning variable c (e.g., text or video) can steer generation across paradigms, but it enters the model at different points. It can be fed to the encoder/decoder or to a conditional prior (VAE, neural-codec+prior), to the prior over discrete tokens (VQ-VAE), as side information in the generator and/or discriminator (GAN), as features or cross-attention inputs in denoisers (diffusion), as additional context in next-step predictors (autoregressive models), or as parameters of base distributions and coupling layers (flows). At sampling time, a single scalar control (e.g., temperature, top-k or top-p sampling) typically trades fidelity for diversity [33]. A related mechanism is *style conditioning*, where a mapping network produces per-layer scales and shifts that modulate activations [34]. We treat style as one more conditioning mechanism, not as a separate generative paradigm.

The *backbones* are the neural function classes that implement the components of a paradigm (encoder, decoder, prior, denoiser, discriminator). They determine key properties such as receptive field, multi-scale paths, context length, and latency. The mapping between paradigms and backbones is many-to-many: the same backbone can instantiate different paradigms, and a single paradigm can be implemented with different backbones. In the next section, we pair each paradigm with typical backbones and highlight the trade-offs that result.

For notation, expectations are $\mathbb{E}[\cdot]$ and the Kullback–Leibler divergence is $D_{\mathrm{KL}}(\cdot\|\cdot)$. We introduce any additional symbols locally in each subsection as needed.

In the next subsections, we survey the generative paradigms considered for sound-effects NAS. For each generative principle we outline: (1) the core learning principle; (2) how conditioning enters; (3) typical backbones; and (4) how sampling and use work in practice. This structure keeps the comparisons compact while emphasizing the control and timing requirements that are specific to sound effects.

B. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [35] learn a generator G_θ that maps noise $z \sim p(z)$ to samples $\hat{x} = G_\theta(z,c)$, while a discriminator D_ψ attempts to distinguish real x from generated $\hat{x}$. The original minimax objective,

$$\min_\theta \max_\psi \; \mathbb{E}_{x\sim p_{\text{data}}}\big[\log D_\psi(x)\big] + \mathbb{E}_{z\sim p(z)}\big[\log\big(1 - D_\psi(G_\theta(z))\big)\big],$$

optimizes a Jensen–Shannon divergence surrogate. In practice, the non-saturating generator loss $\min_\theta \mathbb{E}_z[-\log D_\psi(G_\theta(z))]$ improves gradients [35].

1) Conditioning and Priors: Conditioning variables c can be integrated into the generator G either by concatenating them with the latent vector z or with intermediate feature maps (e.g., [36]). Alternatively, feature modulation can be achieved through style-based or FiLM-like normalization layers that scale and shift activations according to c [37], [38]. On the discriminator side, conditioning is typically enforced via projection or auxiliary heads, allowing D to assess the consistency between the generated signal x and its conditioning input c (e.g., [39]). The latent prior usually follows $z \sim \mathcal{N}(0, I)$. Style-based GANs introduce an additional mapping network that reparameterizes z into a style space, thereby enabling truncation control at inference time [40]. Alternatively, one may learn an implicit conditional prior $p(z \mid c)$ to better align the sampled latents with c without modifying the adversarial objective [41].

2) Backbones: Most adversarial audio architectures adopt convolutional or ResNet-based generators paired with patch or multi-scale discriminators. This design ensures scalable training and stable gradients [36], [39], [42]–[44]. Style-based variants add a mapping network to modulate feature statistics across layers, enabling fine-grained control over timbre and dynamics [38], [40]. Architectures inspired by SinGAN exploit

multi-scale pyramids to capture hierarchical sound structures and facilitate controllable variations [45]. For transient- or impact-dominated signals (e.g., drums), 1D convolutional stacks with multi-scale receptive fields are commonly used [42], [46].

3) Sampling and Use: Inference consists of a single forward pass: sample $z \sim p(z)$ (or $p(z \mid c)$) and compute $\hat{x} = G_\theta(z, c)$. Diversity–quality trade-offs at generation time can be controlled by truncation of z, stochastic noise injection inside G, or rejection of samples with low discriminator scores when available. Continuous latent spaces further allow smooth interpolation and attribute manipulation by traversing meaningful directions in z or in its mapped style space.

C. Variational AutoEncoder (continuous latent)

A Variational AutoEncoder (VAE) [47] assumes that each observation x is generated from a low-dimensional continuous latent z drawn from a prior $p(z)$ and decoded through $p_\theta(x \mid z)$. Because the exact posterior $p_\theta(z \mid x)$ is intractable, an encoder $q_\phi(z \mid x)$ is trained to approximate it while we maximize a tractable lower bound to the log-evidence:

$$\log p_\theta(x) \;\geq\; \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x \mid z)\big] - D_{\mathrm{KL}}\big(q_\phi(z \mid x) \,\|\, p(z)\big).$$

The first term rewards reconstructions that are likely under the decoder; the KL term limits how much information the latent carries and aligns posteriors with the prior so that sampling $z \sim p(z)$ is meaningful. Gradients flow through stochastic sampling via reparameterization [48]:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

making Monte Carlo estimates of the ELBO differentiable in ϕ.

1) Conditioning and priors: When side information c is available, the conditional model replaces $p_\theta(x \mid z)$ by $p_\theta(x \mid z, c)$ and the encoder by $q_\phi(z \mid x, c)$, and often learn $p_\psi(z \mid c)$ to reduce mismatch. Conditioning is commonly injected by concatenation of class tokens at encoder/decoder inputs, as in the conditional Wasserstein autoencoder for drums [49], or by feature-wise modulation (FiLM [37]) to gate intermediate activations in variational pipelines. Learned priors $p_\psi(z \mid c)$ reduce posterior–prior mismatch and make conditional draws coherent. Adversarial matching of the aggregated posterior regularizes z for interpretable control in percussive VAEs [50]. Prior shaping via aligning z to semantic audio embeddings (e.g., CLAP [51]) encourages structured latents that reflect content and improve conditional coherence [52]. Expressive priors implemented as flows over z can further tighten the bound while preserving efficient sampling. Flow-based latent distributions have proven effective for high-fidelity sound-effects synthesis [53].

2) Backbones: Backbone controls how much information routes through z. Convolutional decoders paired with spectrogram likelihoods are reliable (and real-time) in autoencoding settings [49], [54], and feedforward/U-Net-like Convolutional Neural Networks (CNNs) can map high-level controls to waveforms while preserving timbre structure [55]. End-to-end variational systems adapted from text-to-speech add attention blocks and a flow-regularized prior to keep z aligned with semantic factors [52], and phase-aware decoders improve resynthesis fidelity without collapsing the latent space [56].

3) Sampling and use: After training, generation draws $z \sim p(z)$ or $p_\psi(z \mid c)$ and decodes once through $p_\theta(x \mid z, c)$. Interpolations and attribute edits are straightforward in the latent space because z is continuous and usually organized by the information bottleneck. This is especially effective when the backbone and the KL schedule have encouraged z to capture semantically coherent factors of variation.

D. Vector-Quantized Variational Autoencoder (discrete latent)

Vector-Quantized Variational Autoencoders (VQ-VAE) [57] replace the continuous latent by indices from a learned codebook. The encoder $z_e(x)$ is snapped to the nearest code e_k and the decoder predicts $p_\theta(x \mid e_k)$. Training minimizes a reconstruction term plus codebook and commitment losses,

$$\begin{aligned}\mathcal{L}_{\mathrm{VQ}}(x, c) = &\big\|x - \hat{x}\big\|_2^2 \\ &+ \big\|\mathrm{sg}[z_e(x, c)] - e_k\big\|_2^2 \\ &+ \beta \,\big\|z_e(x, c) - \mathrm{sg}[e_k]\big\|_2^2,\end{aligned}$$

and uses a straight-through gradient for the quantizer. Discrete latents ease long-range modeling (via autoregressive or diffusion priors) and support hierarchical or multi-band structure for high fidelity synthesis [58], [59].

1) Conditioning and priors: Side information c can enter the encoder/decoder by *concatenation* (e.g., class or event tags fused into the VQ encoder for conditional sound generation [60]) or by feature-wise modulation that gates intermediate features before quantization (e.g. video-aware conditioning in masked generative setups [61]). The prior over code indices is typically learned: autoregressive Transformers with cross-attention to text or video yield coherent long-form audio from discrete tokens [59], [61], while discrete/latents diffusion refines token sequences or multi-band code streams for higher fidelity and temporal alignment [58], [62].

2) Backbones: Encoders/decoders are usually CNN/ResNet stacks operating on spectrograms or learned codec features, optionally with hierarchical codebooks or multi-rate branches; this aligns well with token priors based on U-Nets or Transformers [58], [59]. Masked modeling backbones (e.g., SpecMaskGIT) also pair naturally with discrete latents for efficient pretraining and controllable synthesis from sparse context [63]. For video-to-audio (VTA), dual-stream designs keep a lightweight VQ decoder while the prior backbone (Transformer or dual-U-Net) handles alignment and rhythm cues [61].

3) Sampling and use: Generation selects code indices from the learned prior conditioned on c and decodes once through the VQ decoder. Hierarchical or multi-band codebooks enable coarse-to-fine control and robust editing (token replacement or span infilling) [58]. In text/video-to-audio, discrete-token priors improve long-horizon structure and synchronization while keeping inference tractable in practice [59], [61].

E. Normalizing Flows

Normalizing flows (NFs) [48] construct an invertible map $f_\theta : \mathcal{X} \to \mathcal{Z}$ that transforms data x to a simple base latent representation $z = f_\theta(x)$ with tractable density $p(z)$ (e.g., standard Gaussian). By change of variables,

$$\log p_\theta(x \mid c) = \log p_\epsilon\big(f_\theta(x,c) \mid c\big) + \log\left|\det \frac{\partial f_\theta(x,c)}{\partial x}\right|,$$

so maximum likelihood is exact and trained by gradient ascent on $\sum_i \log p_\theta(x^{(i)})$. Expressivity comes from composing K simple, bijective layers $f_\theta = f_K \circ \cdots \circ f_1$ for which both the inverse and $\log|\det J|$ are efficient.

1) Conditioning and priors: Conditioning enters the affine coupling networks by concatenation of c (class/action/timing) into the scale/shift subnets, enabling controllable sound effects synthesis [64]. In flow-matching variants, cross-attention to text embeddings provides semantic guidance during the learned transport [65], while video features (and alignment cues) guide the trajectory for efficient VTA generation under rectified/flow matching [66]. Learned conditional bases replace a fixed p_ϵ with $p_\epsilon(\cdot \mid c)$, improving coherence without altering invertibility [67].

2) Backbones: Flow coupling networks are typically CNN/ResNet stacks (with invertible 1×1 convs) operating on log-magnitude spectrograms or learned features [64]. Flow-matching TTA/Video-to-Audio (V2A) systems pair the transport field with U-Net or Transformer blocks to fuse text/video context while keeping the sampler lightweight [65], [67]. Latent-flow designs move the transport to a compact space, easing long-horizon structure while the decoder handles fine detail [67].

3) Sampling and use: Exact flows sample in a single forward pass through the invertible stack, giving real-time or near–real-time synthesis and smooth control sweeps for sound effects parameters [64]. Flow matching solves a short ODE (or rectified flow) with tens of steps, trading tiny likelihood slack for speed while preserving semantic alignment from text/video [65], [66]. In practice, flows excel when low-latency generation and deterministic, repeatable control are required, and when conditioning carries precise timing.

F. Diffusion

Diffusion models learn to reverse a corruption process that gradually mixes data with noise. In discrete-time Denoising Diffusion Probability Models (DDPMs) [68], the forward process is

$$x_t = \sqrt{\bar\alpha_t}\, x_0 + \sqrt{1-\bar\alpha_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

and the model predicts either the noise $\epsilon_\theta(x_t, t)$, the clean signal $x_{0,\theta}(x_t, t)$; the common training loss is

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{x_0,\, c,\, \epsilon,\, t}\Big[\, \big\|\epsilon - \epsilon_\theta(x_t, t, c)\big\|_2^2 \,\Big],$$

which is a form of denoising score matching. In continuous time, the same idea appears as an SDE with a learned score $s_\theta(x,t) \approx \nabla_x \log p_t(x)$ and a corresponding probability-flow ODE for deterministic generation.

1) Conditioning and priors: Conditioning is typically injected by cross-attention to text/video encodings, letting the denoiser network (usually Diffusion Transformers or U-Net-style CNN) attend to semantic and temporal cues throughout the denoising trajectory [69]. Precise control signals (e.g., onsets, timestamps, pitch tracks) can be fed by simple concatenation to inputs or intermediate blocks for fine timing and event placement [53]. When stronger gating is needed, FiLM/style modulation scales and shifts residual features using conditioning embeddings, supporting multi-condition fusion without overfitting [70]. For long-range structure and fidelity, discrete codec-token priors provide a robust target space in which diffusion operates (often multi-band), improving coherence and sample quality [58].

2) Backbones: We found many systems in the sound effects NAS literature that use U-Net backbones over spectrograms or latents with multi-scale residual blocks and attention [5], [71], [72]. Dual-U-Net designs decouple content and timing (or text and audio latents) to boost sync and detail [73], [74]. In parallel, Diffusion Transformers (DiTs) have become the prevailing backbone in many large-scale diffusion models across domains, and Transformer-based denoisers are increasingly adopted for audio and audiovisual generation as well. Transformer-augmented or temporal-aware diffusion improves long-horizon coherence [75], [76], and specialized backbones target precise event control (e.g., rhythm/onset modules or control signals) [53], [70]. Waveform-domain diffusion trades speed for phase fidelity [77], [78].

3) Sampling and use: Sampling iteratively denoises in T steps (spectrogram or latent) with classifier-free guidance. Few-step paths use consistency distillation or latent-consistency to cut steps to single-digits while keeping quality [79], [80]. For TTA, latent sampling with cross-attended prompts followed by vocoder decoding is standard in open-domain pipelines [76]. VTA uses the same sampler but conditions on per-frame features to lock temporal sync [69], [81]. Editing and insertion rely on masked/instruction-guided diffusion to replace spans while preserving context [82]. When precise events are needed, segment-wise resampling or timing-conditioned latents enforce timestamps and reduce drift [53].

G. Autoregressive

Autoregressive (AR) models factorize the data distribution into next–step conditionals,

$$p_\theta(x_{1:T} \mid c) = \prod_{t=1}^{T} p_\theta\big(x_t \mid x_{<t},\, c\big),$$

and typically train by maximum likelihood with teacher forcing,

$$\mathcal{L}_{\text{AR}}(\theta) = -\,\mathbb{E}_{x,\,c} \sum_{t=1}^{T} \log p_\theta\big(x_t \mid x_{<t},\, c\big).$$

This objective is exact under the AR factorization, but it is local: it optimizes one-step prediction and does not impose a global transport/energy objective nor a latent bottleneck. Generative behavior emerges at inference by chaining predictions.

1) Conditioning and priors: Conditioning c is injected as additional context to the next-token predictor via cross-attention (TTA in discrete-token AR transformers [59], VTA with masked/AR fusion [61]) or simple concatenation of class/scene cues in conditional AR RNNs [83]. There is no separate latent prior $p(z)$: the autoregressive model itself defines $p_\theta(x_t \mid x_{<t}, c)$ over tokens. Masked-token variants learn an inpainting prior that fills spans conditioned on context and c [63].

2) Backbones: The paradigm (next-step likelihood with causal context) is agnostic to the backbone. RNN/GRU/LSTM [84], [85] stacks provide streaming with compact state [83], [86]. Causal Transformers scale context and support rich cross-modal conditioning for discrete codec/VQ tokens [59], [87]. Masked transformers (BERT-style over spectrogram/code tokens) enable efficient pretraining and controllable span infilling for VAT/TTA [61], [63].

3) Sampling and use: Generation proceeds sequentially from a prompt (or special token), sampling $x_t \sim p_\theta(\cdot \mid x_{<t}, c)$; temperature and top-$k$/$p$ govern the fidelity–diversity trade-off, while KV/state caching keeps throughput workable for long sequences [87]. TTA uses cross-attended prompts over discrete tokens and decodes with the codec's vocoder [59]; VTA adds frame-level context to maintain synchronicity [61]. For editing and insertion, masked AR sampling fills selected spans conditioned on surrounding tokens and controls [63].

H. Neural-codec + Prior

Rather than introducing a new generative principle, this subsection groups models that separate representation from generation. A neural codec (e.g. EnCodec [88]) learns an encoder–decoder pair (E_ϕ, D_θ) that maps x onto z (a low-rate latent stream) and back $\hat{x} = D_\theta(z)$, optimized with a rate–distortion objective

$$\min_{\phi,\theta} \; \mathbb{E}_{x,\,c}\Big[d\big(x,\; D_\theta(E_\phi(x,c),\,c)\big) \;+\; \lambda\, R\big(E_\phi(x,c) \mid c\big)\Big],$$

where $d(\cdot,\cdot)$ measures reconstruction error and $R(\cdot)$ penalizes bitrate (e.g., entropy of discrete tokens). After the codec is trained, a prior is learned on the latent sequence z using one of the paradigms described in previous subsections (typically an autoregressive or diffusion model). We treat this configuration separately for organizational reasons, because the codec fixes a bandwidth-limited, perceptually shaped space and decouples compression from generation, while the prior supplies semantic structure and diversity.

In addition to proper neural codecs, some works use self-supervised audio encoders such as wav2vec 2.0 [89] as sources of continuous (and, when applicable, quantized) latent representations. Although these models are not trained as codecs—they do not optimize a waveform reconstruction objective—they can play a similar role in that a generative prior can operate directly on their learned latents. Such Self-Supervised Learning (SSL) representations have been shown to transfer well across downstream tasks, and analyses suggest that different hidden layers capture complementary information at different levels of abstraction [90].

1) Conditioning and priors: Conditioning c is injected mainly into the prior. For discrete tokens, text is fused through cross-attention in an AR Transformer prior (TTA with EnCodec/RVQ tokens) [59], and video cues are integrated via masked/AR fusion for V2A alignment [61]. Diffusion priors operate directly in token/latent space, using classifier-free guidance and cross-attended embeddings for fidelity and long-range structure [58], [91]. Discrete diffusion over code indices offers robust coarse control [62]. Retrieval-augmented conditioning further stabilizes semantics by biasing the prior with nearest exemplars [92].

2) Backbones: Codecs are lightweight CNN/ResNet stacks with down/up paths and residual vector quantization to yield low-rate, streaming-friendly tokens. These pair naturally with causal Transformer priors for long contexts in TTA [59], [87]. When priors are diffusion models, U-Net/Transformer backbones denoise in the token/latent domain, and dual-U-Net designs help separate content versus timing factors [58], [73]. Masked modeling backbones also work well with discrete latents for efficient pretraining and controllable span infilling, especially in V2A [61], [63].

3) Sampling and use: Synthesis is two-stage: (1) sample $\tilde{z} \sim p_\psi(z \mid c)$ with AR (temperature/top-k/p) or diffusion (step budget/guidance), then (2) decode once $\hat{x} = D_\theta(\tilde{z})$. AR priors excel at semantic coherence and prompt following over long horizons [59], [87], while diffusion priors provide strong fidelity and editable generations via masked spans or token-level replacement [58], [63]. Hierarchical Residual Vector Quantization (RVQ) enables coarse-to-fine sampling and localized edits (modify only high-level bands/tokens) [58].

I. DDSP-hybrids

DDSP denotes a family of methods that embed differentiable DSP modules (oscillators, filters, controllers) in the graph so a network predicts interpretable synthesis parameters and audio is rendered by a differentiable synthesizer or signal-processing module. Let ψ be parameter trajectories (e.g., f_0, amplitudes, envelopes, filter coefficients) and g_θ the differentiable synthesizer. Training is based on the analysis-by-synthesis paradigm:

$$\hat{\psi} = h_\phi(x,c), \quad \hat{x} = g_\theta(\hat{\psi}), \quad \min_{\phi,\theta} \mathbb{E}\big[d(x,\hat{x}) + \textstyle\sum_j \lambda_j\, \mathcal{R}_j(\hat{\psi})\big],$$

with perceptual $d(\cdot,\cdot)$ and regularizers $\mathcal{R}_j$ (smoothness, non-negativity, bandwidth). This family does not follow a generative objective per se. It supplies a strong inductive bias (harmonicity, stability, controllability) and delegates generativity to the model that produces or modulates ψ.

1) Conditioning and priors: Conditioning c (text, score, identity, control curves) can drive ψ directly—supervising trajectories such as spectral envelopes for sound effects controls [93]—or via a learned prior $p_\psi(\psi \mid c)$ that samples plausible parameter paths before rendering. For generative behavior, a sequence prior over ψ (AR or diffusion in parameter space) provides diversity while constraints enforce physical ranges and smoothness. This yields interpretable, editable controls (e.g., transient vs. sustain shaping in percussives) [94], [95].

2) Backbones: The family is the differentiable DSP decoder g_θ; backbones are the networks mapping $(x, c) \to \psi$ and the priors over ψ. Compact CNN/ResNet stacks give stable, low-latency parameter estimates (timbre and envelope controls) [55]. When long-range structure matters (phrasing, form), lightweight Transformer blocks can be added without bypassing ψ semantics. Filterbank-based controllers provide robust time-varying trajectories for noisy sound effects and pair well with DDSP renderers [96]. Small residual neural decoders can capture off-manifold details while keeping most energy in interpretable paths [93].

3) Sampling and use: Synthesis is two-stage: (1) obtain or sample $\tilde{\psi} \sim p_\psi(\psi \mid c)$ with AR/diffusion and constraint checks; (2) render once $\hat{x} = g_\theta(\tilde{\psi})$. It supports real-time operation, direct edits to pitch/brightness/articulation, and robust generalization where DSP assumptions hold (e.g., tonal/percussive events, vocalizations) [94], [97]. Hybrids mitigate expressivity limits with hierarchical controls and modest neural residuals, preserving the interpretability that motivates DDSP [93].

VI. Overview of Neural Sound-Effects Synthesis

Table I compiles published contributions organized by Farnell's sound-effect categories (rows) and by generative paradigm (columns).

It's worth noting that natural and lifeform-related sound effects significantly outnumber science fiction and Mayhem-related ones. This can be attributed to the abundance of these sound effects in nature, which are readily available in large, open online databases like YouTube. This natural prevalence provides a wealth of data for training models and creates a higher demand for replicating these sounds. Additionally, certain sound effects, particularly percussive sounds, whether from acoustic drums or human activities (e.g., footsteps), seem to pique the community's interest more than others.

Furthermore, it is uncommon for articles to focus solely on producing one type of sound effect. Instead, it is typical to find references used for multiple effects, likely driven by two factors: pursuing innovative sound combinations that do not naturally occur and including a diverse range of effects within the datasets. The idea of a generative model dedicated to a single purpose might seem unusual, but this largely depends on the level of control afforded to the user. Models trained on a wide variety of effects may become overly complex if they lack sufficient control mechanisms. This complexity is one of the reasons behind the development of multimodal systems, which aim to integrate and manage different types of effects more effectively.

Table II outlines all the articles selected for this scope review, categorized according to the architecture employed and the dataset used for training. Unless explicitly stated otherwise, it is assumed that the architecture can generate all elements within the respective dataset.

It is noteworthy how certain datasets have become particularly influential in the field, with AudioSet [98] being a prominent example and, more recently, AudioCaps [99], especially in the context of multimodal research. Many of the analyzed datasets incorporate multimodal data along with audio, such as video or text, which enriches the training data. It is also common to find that some datasets are actually subsets of larger datasets, a trend especially evident with AudioSet. Another notable observation is the number of studies that use privately curated datasets or choose not to disclose the specific training set employed. This trend suggests a degree of proprietary research or the desire to protect competitive advantages. In contrast, many other datasets are sourced from open repositories donated by the community, with Freesound [100] being a particularly successful example.

Interestingly, with the exception of the BBC SFX [101] and some others, few datasets are entirely focused on sound effects, often necessitating significant curation before they can be effectively utilized for this specific task. This need for curation underscores the challenges and the utmost importance of careful dataset selection and preprocessing in achieving high-quality sound effect generation.

A quantitative analysis of Table I reveals the underlying distribution in research scope across sound taxonomies, measured by the number of sound effect entries citing relevant work within the table. The "Nature" taxonomy shows the highest engagement with 34 unique entries, followed by "Lifeforms" (32 unique entries) and "Idiophonics" (24 unique entries). "Mayhem" (18 unique entries), "Machinery" (17 unique entries), and "Artificial" (14 unique entries) are represented, while "Sci-Fi" (1 unique entry) is addressed least frequently according to this table's structure. Regarding paradigm prevalence, summing the citation instances across both Table I and Table II provides a comprehensive picture. Diffusion models appear most frequently with a total of 67 unique citations, followed by GAN with 36 unique citations. VAE, VQ-VAE, Neural-codec, and Autoregresive are also prominent, featuring in 18, 14, 15, and 20 unique citations each respectively. Normalizing flows and DDSP hybrids are less representative, featuring 5 unique citations each of them.

VII. Multimodality in NAS

Multimodality in generative audio has accelerated as a practical response to the difficulty of specifying exactly the sound a user wants to synthesize. Describing a sound effect with all its nuances is hard. To bridge this gap, systems condition on text$^{\text{T}}$, images$^{\text{I}}$, and video$^{\text{V}}$ (see Tables I–II for the consolidated references). Text-driven approaches$^{\text{T}}$ capture semantics and style; image–audio methods$^{\text{I}}$ link visual context to acoustic events; and VTA systems$^{\text{V}}$ address the added challenge of temporal alignment and synchronization. Some works even explore onomatopoeia as a compact, controllable textual proxy for sound [102]–[104].

As shown in Fig. 5, multimodal systems have rapidly evolved: since 2023 they match unimodal approaches in sound-effect generation, and by 2024 they clearly surpass them.

Despite this momentum, most multimodal pipelines still translate between modalities rather than learn a shared representation. Unified latent spaces that jointly encode audio and vision remain rare. Instead, information commonly flows from the visual domain to audio, often with text as an

intermediate control signal. This has direct implications for controllability—a persistent challenge in NAS—where richer, truly joint-representations of modes would likely yield finer and more reliable control.

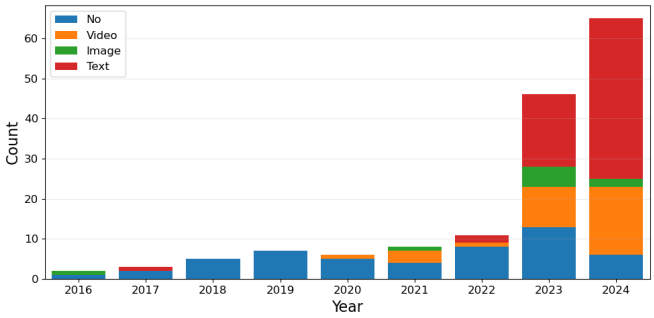

Fig. 5. Number of articles using (or not) multimodality over the years.

VIII. Evaluation Methods

Progress in sound effects synthesis depends also on sound evaluation. Here we emphasize audio-quality metrics—fidelity and listener perception. We describe the metrics for evaluating generated audio quality in terms of fidelity and perceived quality. Metrics used exclusively to assess any other type of quality (e.g., image, text, in multimodal systems) are out of our scope. Table III shows papers following each of the evaluation methods considered, while Fig. 6 shows the relative distribution of the evaluation methods.

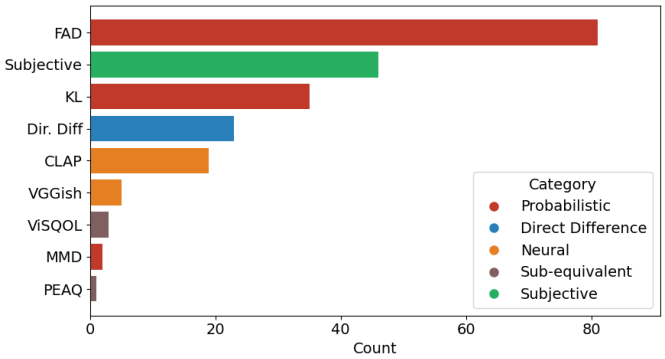

Fig. 6. Most popular evaluation methods found in the reviewed literature.

A. Objective Methods

In audio evaluation, many objective metrics are employed to assess the performance of generative models and other audio processing algorithms. These metrics quantify various aspects of audio quality, such as spectral coherence, reconstruction fidelity, or adaptation to certain sound effects. They ensure a comprehensive evaluation and are necessary for comparison with other state-of-the-art techniques. However, they do not necessarily match with subjective quality. Below, we describe the metrics used in the articles considered in this document:

Direct difference

It measures the distance between the generated and true values. They are usually used to train the models. These metrics are not specifically related to generative models since the comparison needs to be done against a real signal. Generative models are not meant to recreate exactly what can be found in the training datasets, but a plausible related sound effect. Mean Squared Error (MSE) and Mean Absolute Error (MAE) are usually computed for spectral distance provided by the Short-Time Fourier Transform (STFT), or the MEL representation. Some authors used a multi-scale representation in this regard. In the case of [105] they used a MEL Cepstral Distortion (MCD) as their direct comparison, and in [94] the Spectral Flux (SF).

Statistical Difference

It computes the statistical distance of certain audio characteristics between the generated sound and a set of sounds belonging to the same category. These metrics are suitable for generative model evaluation as they prove the sound effects contain the characteristics of those representing the same sound effects, considering the expected inherent variability found in the true values. These metrics are usually computed in the model's latent space z, as it contains fundamental information on the audio features. Some of the metrics found in the literature are:

- Kullback-Leibler Divergence (KLD) [106]: Relies on entropy to measure how one probability distribution diverges from a reference distribution. In the context of audio evaluation, it quantifies the difference in information content (or entropy) between the original and processed audio signals.
- Fréchet Audio Distance (FAD) [107]: A metric that assesses the similarity between the statistical properties of features from original and processed audio signals, specifically by comparing their means and covariances in a multidimensional feature space. Unlike the KLD, FAD offers a symmetric evaluation of the overall distributional difference, providing a more comprehensive assessment of how closely the processed audio resembles the original while being less sensitive to small discrepancies or outliers in the data.

Neural Evaluation

These metrics use a neural model to evaluate the quality of the generated sound effect. Two models are popular in the literature:

- VGGish Loss [108]: Pre-trained deep neural network model designed for audio classification tasks. It is based on the VGG architecture and has been adapted to work with audio data by processing log-mel spectrograms. VGGish is widely used for feature extraction in audio tasks, converting raw audio inputs into compact, high-level feature embeddings that can be used for audio similarity assessment. The model is trained on a large-scale dataset, which enables it to generalize well across diverse audio domains.

- Contrastive Language-Audio Pretraining (CLAP) [109]: CLAP is a model designed for learning joint multimodal representations of audio and text through contrastive learning. By aligning audio and textual embeddings in a shared latent space, CLAP facilitates tasks such as cross-modal retrieval and classification. The model uses a contrastive loss function to ensure that paired audio-text samples are closely aligned while separating non-paired samples. It is particularly effective for applications that require understanding the relationship between audio content and corresponding textual descriptions.

Subjective-equivalent Evaluation

These are objective metrics that claim to mimic human perception in terms of quality evaluation. They compare the generated sample with an original recording, mapping the difference to a subjective scale. They can be taken as a proxy for the opinion of final users. The metrics considered are:

- Virtual Speech Quality Objective Listener (ViSQOL) [110]: A perceptual model used to objectively assess the quality of speech and audio signals by simulating human auditory perception. It compares the time-frequency content of a processed audio signal to a reference signal using spectrograms and a similarity measure based on the Structural Similarity Index (SSIM). It gives a ViSQOL-MOS metric with values between 1 (worst) and 5 (best). The MOS-ViSQOL mapper can be fine-tuned with human subjective test data to improve reliability and domain match, as in [111].
- Perceptual Evaluation of Audio Quality (PEAQ) [112]: An objective metric designed to assess the perceived audio quality of processed signals, particularly in the context of audio compression and transmission. It simulates human auditory perception by analyzing the differences between a reference signal and a test signal across various perceptual domains, such as frequency, loudness, and temporal masking.

B. Subjective Methods

Subjective audio evaluation methods are the preferred method for assessing the performance and quality of generative audio models. Unlike objective metrics, subjective evaluations rely on human perception to determine audio quality. This approach allows testing end-users, which may include more than just the quality of the recording but also the application, the environment, or the psychology of the use case. As such, subjective metrics provide insights that objective measures may overlook, capturing subtle aspects of auditory perception essential for a comprehensive evaluation.

Several factors must be considered when conducting a subjective evaluation to ensure reliability and validity. The testing environment should be acoustically controlled to minimize external noise and distractions. Participants should have normal hearing and be adequately trained to understand the evaluation criteria. Randomizing the presentation of audio samples helps prevent order effects, and a sufficiently large sample size ensures that the results are statistically significant. However, a definitive standard process does not exist to assess the perceived quality of generative sound. Therefore, one may find creative ways to analyze the generative model performance. Here, we will focus on the main premises and cite the articles that performed the subjective tests without going deeper into the details.

One of the most widely used subjective evaluation metrics is the Mean Opinion Score (MOS) [113]. MOS is a scalar measure used to rate the quality of audio on a scale from 1 (bad) to 5 (excellent). In a typical MOS test, a group of listeners rate the quality of various audio samples, and the scores are averaged to produce the MOS. This metric is popular due to its simplicity and effectiveness in capturing listener preferences and perceptions.

The MUlti-Stimulus test with Hidden Reference and Anchor (MUSHRA) [114] is a robust method for subjective audio evaluation, originally designed to assess intermediate quality levels of audio codecs. Specifically, it employs simultaneous comparison of multiple stimuli with reference and anchor signals, facilitating detailed and discriminative assessments of audio quality in a manner conceptually similar to the standard MUSHRA approach.

IX. Discussion

This scope review consolidates the rapidly evolving landscape of Neural Sound Effect Synthesis, mapping the key architectures, datasets, sound taxonomies, and evaluation strategies reported in the literature. The breadth and dynamism of the field hinder the formulation of a prescriptive "cookbook" to approach the neural synthesis—a simple mapping from a desired sound effect to a definitive paradigm and dataset—currently is unclear. The structured analysis presented here offers insights and navigational guidance for different stakeholders within the sound synthesis community. The trends observed, particularly in generative paradigm adoption and conditioning strategies, alongside the detailed cross-referencing in Tables I and II, may illuminate both the current state and potential future directions.

A prominent trend, vividly illustrated in Figure 4, is the temporal evolution of paradigm preference. While earlier work leveraged GANs, VAEs, and Autoregressive principles to establish foundational sound effect synthesis capabilities, the recent surge (post-2022) in Diffusion models indicates a paradigm shift. These paradigms appear to offer superior performance in capturing the complexity and temporal dependencies inherent in many sound effects, driving much of the contemporary research focus. This architectural shift correlates strongly with the increasing sophistication behind conditioning methods, as shown in Figure 5. The move from unconditional generation towards text, image, and video conditioning reflects a growing demand for controllability, contextual relevance, and multimodality, pushing Neural Synthesis beyond simple sound-effect generation towards more integrated creative tools. In addressing the relationship between generative paradigms and sound taxonomies (Table I), a definitive one-to-one mapping remains ineffable. While one might hypothesize that certain generative principles are inherently better suited to specific sound classes (e.g., transient versus textural sounds),

TABLE I

SUMMARY OF THE ARTICLES WHERE THE SYNTHESIS OF A SPECIFIC SOUND EFFECT IS MENTIONED. SUPERSCRIPTS INDICATE MULTIMODAL CONDITIONING (V FOR VIDEO, I FOR IMAGE AND T FOR TEXT).

NAS SUMMARY OVER TAXONOMY									
		PRINCIPLE							
TAXONOMY	SOUND	GAN	VAE	VQ-VAE	Normalizing Flow	Diffusion	Autoregresive	Neural-Codec Prior	DDSP
Artificial	Car							[115]^{V}, [116]^{I}	
	Clock							[115]^{V}	
	Keyboard	[117]^{T}, [118]^{I}, [119], [120]	[118]^{I}, [73]^{T}, [121], [122]^{I,T}	[123]^{T}, [73]^{T}, [122]^{I,T}	[118]^{I}	[77]^{T}, [117]^{T}, [73]^{T}, [121], [124], [122]^{I,T}, [125]	[123]^{T}, [126]^{I}, [122]^{I,T}		
	Typing							[115]^{V}	
Idiophonics	Beats		[56]						
	Break							[115]^{V}	
	Cutting							[115]^{V}	
	Cymbals	[46]							[94]
	Drums	[4], [127], [46], [40], [38], [42], [128], [95] [50]	[49], [50]			[78]		[55]	[94]
	Hits	[41]		[129]^{I}			[129]^{I}		
	Impact		[130]			[131], [132]^{V}, [133]^{T,V}			
	Jump	[45]							
	Kick	[46]							
	Knocking	[36]	[56]						[96]
	Metal								[96]
	Pottery								
	Snare	[46]							
Lifeforms	Animals	[134]^{T}	[134]^{T}			[134]^{T}		[116]^{I}	
	Applause						[135]	[136]	
	Babble						[86]		
	Bees						[135]	[136]	
	Birds	[39], [137]^{V}, [127], [138]^{V}				[137]^{V}	[135]	[136]	[97]
	Cats								[97]
	Cough	[117]^{T}, [118]^{I}, [120]	[118]^{I}, [73]^{T}, [122]^{I,T}	[123]^{T}, [73]^{T}, [122]^{I,T}	[118]^{I}	[77]^{T}, [117]^{T}, [73]^{T}, [124], [122]^{I,T}, [125]	[123]^{T}, [126]^{I}, [122]^{I,T}		
	Cows								[97]
	Crickets						[135]	[136]	
	Crowd						[135]	[136]	
	Crows								[97]
	Dog bark	[117]^{T}, [118]^{I}, [139]^{V}, [119], [120]	[118]^{I}, [73]^{T}, [121], [122]^{I,T}	[123]^{T}, [73]^{T}, [122]^{I,T}	[118]^{I}	[77]^{T}, [117]^{T}, [73]^{T}, [121], [124], [122]^{I,T}, [125]	[123]^{T}, [126]^{I}, [122]^{I,T}		[97]
	Eating sounds	[140]^{V}							
	Footsteps	[45], [39], [43], [117]^{T}, [118]^{I}, [41], [119], [120]	[118]^{I}, [93], [73]^{T}, [121], [141], [122]^{I,T}	[129]^{I}, [123]^{T}, [73]^{T}, [122]^{I,T}	[118]^{I}	[77]^{T}, [117]^{T}, [73]^{T}, [121], [124], [122]^{I,T}, [125]	[129]^{I}, [123]^{T}, [126]^{I}, [122]^{I,T}	[115]^{V}	[96], [93]
	Frogs								[97]
	Hens								[97]
	Horse							[115]^{V}	
	Insects	[137]^{V}, [138]^{V}				[137]^{V}			[97]
	Pigs								[97]
	Sheep								[97]
	Sneeze	[117]^{T}, [118]^{I}, [119], [120]	[118]^{I}, [73]^{T}, [121], [122]^{I,T}	[123]^{T}, [73]^{T}, [122]^{I,T}	[118]^{I}	[77]^{T}, [117]^{T}, [73]^{T}, [121], [124], [122]^{I,T}, [125]	[123]^{T}, [126]^{I}, [122]^{I,T}		
Machinery	Machine	[134]^{T}	[134]^{T}			[134]^{T}	[135]		
	Electromagnetism		[142]						
	Engine	[39]							
	Motor	[117]^{T}, [118]^{I}, [119], [120]	[118]^{I}, [93], [73]^{T}, [121], [122]^{I,T}	[123]^{T}, [73]^{T}, [122]^{I,T}	[118]^{I}	[77]^{T}, [117]^{T}, [73]^{T}, [121], [124], [122]^{I,T}, [125]	[123]^{T}, [126]^{I}, [122]^{I,T}		[93]
Mayhem	Explosion				[64]^{T}				
	Gunshot	[39], [117]^{T}, [118]^{I}, [41], [119], [120]	[118]^{I}, [93], [73]^{T}, [121], [122]^{I,T}	[129]^{I}, [123]^{T}, [73]^{T}, [122]^{I,T}	[118]^{I}	[77]^{T}, [117]^{T}, [73]^{T}, [121], [124], [122]^{I,T}, [125]	[129]^{I}, [123]^{T}, [126]^{I}, [122]^{I,T}	[115]^{V}	[93]
Nature	Environment	[138]^{V}, [143], [144]^{T}, [44]		[144]^{T}, [145], [58]^{V}		[91]^{T}, [131], [79]^{T}, [58]^{V}	[83], [103]^{V}, [146]^{T}	[146]^{T}, [116]^{I}	[145]
	Fire						[135]	[136], [115]^{V}	
	Fire Crackles	[147]							
	Fire Rumbles	[147]							
	Nature	[134]^{T}	[134]^{T}			[134]^{T}	[148]^{T}		
	Rain	[39], [117]^{T}, [118]^{I}, [119], [120]	[118]^{I}, [73]^{T}, [121], [122]^{I,T}	[123]^{T}, [73]^{T}, [122]^{I,T}	[118]^{I}	[77]^{T}, [117]^{T}, [73]^{T}, [121], [124], [122]^{I,T}, [125]	[123]^{T}, [126]^{I}, [122]^{I,T}	[136], [115]^{V}	
	Sink							[136]	
	Thunder							[115]^{V}	[96]
	Water		[142]				[135]		
	Waterfall							[115]^{V}	
	Wind						[135]	[136]	
Sci-Fi	Light Saber		[142]						

IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

TABLE II

SUMMARY OF THE STATE OF THE ART WITH RESPECT TO THE DATASET USED FOR TRAINING. ARTICLES THAT APPEARED IN TABLE I ARE IN BOLD. ARTICLES THAT USED A PRIVATE DATASET OR DID NOT EXPLICITLY MENTION ANY DATASET APPEAR UNDER THE KEY "PRIVATE" AND "UNSPECIFIED," RESPECTIVELY. SUPERSCRIPTS INDICATE MULTIMODAL CONDITIONING.

NAS SUMMARY OVER DATASET								
	PRINCIPLE							
DATASET	GAN	VAE	VQ-VAE	Normalizing Flow	Diffusion	Autoregressive	Neural-Codec Prior	DDSP
200 Drum machines	**[4]**							
Adobe Audition SFX					[149]^{T}			
AFD	**[138]**^{V}						**[115]**^{V}	
Audio-Alpaca					[150]^{T}			
AudioCaps [99]	**[144]**^{T}, **[134]**^{T}	[151]^{T}, [65]^{T}, [152]^{T}, **[134]**^{T}	[63]^{T}, [59]^{T}, **[144]**^{T}, [62]^{T}	[67]^{T}	[29]$^{\dot{T},V}$, [63]^{T}, [153]$^{\dot{T}}$, [154]$^{\dot{T}}$, [151]$^{\dot{T}}$, [149]$^{\dot{T}}$, [155]$^{I,\dot{T},V}$, [65]$^{\dot{T}}$ [156]^{T}, [80]^{T}, [72]^{T}, [5]^{T}, [157]^{T,V}, [158]^{I,T}, [159]^{T,V}, **[91]**^{T} **[79]**^{T}, [87], [160]^{T}, [82]^{T}, [152]^{T}, [161]^{T}, [162]^{T}, [163]^{T,V} [62]^{T}, **[134]**^{T}, [92]^{T}, [164]^{T}	[63]^{T}, [59]^{T}, [62]^{T}	[165]^{T}, [105]^{T}	
AudioSet [98]	**[137]**^{V}, **[127]**, **[117]**^{T}, [166], **[134]**^{T}	[151]^{T}, [52]^{T}, **[134]**^{T}	[63]^{T}, [59]^{T}, **[58]**^{V}, [62]^{T}	[67]^{T}	[167]^{V}, **[137]**^{V}, [63]$^{\dot{T}}$, [70]$^{\dot{T}}$, [151]$^{\dot{T}}$, [168]$^{\dot{V}}$, [149]$^{\dot{T}}$, [155]$^{I,\dot{T},V}$ **[117]**^{T}, [52]^{T}, [169]$^{T,\dot{V}}$, [5]^{T}, [69]$^{\dot{V}}$, [76]^{T}, **[58]**$^{\dot{V}}$, [87] [82]^{T}, [161]^{T}, [162]^{T}, [163]$^{T,\dot{V}}$, [62]^{T}, **[134]**^{T}, [164]^{T}	[167]$^{\dot{V}}$, [63]^{T}, [59]^{T}, [62]^{T}	[105]^{T}, [81]$^{T,\dot{V}}$	
AudioSparx		[53]^{T}			[53]$^{\dot{T}}$			
Audiostock					[149]$^{\dot{T}}$, [155]$^{I,\dot{T},V}$			
AVSync15 [170]					[168]$^{\dot{V}}$		[81]$^{T,\dot{V}}$	
BBC SFX	**[39]**, **[118]**^{I}, **[41]**	[151]$^{\dot{T}}$, [52]^{T}, **[118]**^{I}	[59]$^{\dot{T}}$, **[129]**^{I}, **[123]**^{T}	[67]$^{\dot{T}}$, **[118]**^{I}	[171], [151]$^{\dot{T}}$, [155]$^{I,\dot{T},V}$, [52]$^{\dot{T}}$, [169]$^{\dot{T},V}$, [5]$^{\dot{T}}$, **[124]**, [161]$^{\dot{T}}$	[59]$^{\dot{T}}$, **[129]**^{I}, **[123]**^{T}, **[126]**^{I}		
Boom Library	**[39]**, **[41]**		**[129]**^{I}			**[129]**^{I}		
Clotho [172]	**[117]**$^{\dot{T}}$, **[144]**$^{\dot{T}}$	[151]$^{\dot{T}}$, [152]$^{\dot{T}}$	[59]$^{\dot{T}}$, **[144]**$^{\dot{T}}$		[151]$^{\dot{T}}$, [149]$^{\dot{T}}$, **[117]**$^{\dot{T}}$, [152]$^{\dot{T}}$, [161]$^{\dot{T}}$	[59]$^{\dot{T}}$	[165]$^{\dot{T}}$	
CountixAV [173]						[174]$^{\dot{V}}$		
DCASE [175]	**[41]**, [176]	[52]$^{\dot{T}}$, **[73]**$^{\dot{T}}$, **[121]**, **[122]**$^{I,\dot{T}}$	**[73]**$^{\dot{T}}$, **[122]**$^{I,\dot{T}}$		**[77]**$^{\dot{T}}$, [171], [52]$^{\dot{T}}$, **[73]**$^{\dot{T}}$, **[121]**, **[122]**$^{I,\dot{T}}$, [164]$^{\dot{T}}$, **[125]** [71]	**[83]**, **[122]**$^{I,\dot{T}}$		
Ego4D [177]					[178]$^{\dot{V}}$			
ENST-Drums	**[38]**							
EPIC-KITCHENS [179]					[178]$^{\dot{V}}$			
Epidemic Sound					[149]$^{\dot{T}}$, [155]$^{I,\dot{T},V}$			
ESC [180]	**[127]**, **[44]**		**[145]**		[149]$^{\dot{T}}$, [155]$^{I,\dot{T},V}$, [156]$^{\dot{T}}$, **[91]**$^{\dot{T}}$, [82]$^{\dot{T}}$, [161]$^{\dot{T}}$			**[97]**, **[145]**
Freesound [100]	**[45]**, **[147]**, **[134]**^{T}	[151]$^{\dot{T}}$, [52]^{T}, **[134]**^{T}		[67]^{T}	[151]$^{\dot{T}}$, [52]^{T}, [156]^{T}, [5]^{T}, [161]^{T}, **[134]**^{T}, [164]^{T}		**[55]**	**[96]**
FreeToUseSounds	**[117]**^{T}		[59]^{T}		[155]^{I,T,V}, **[117]**^{T}	[59]^{T}		
FSD50K [181]	**[118]**^{I}	**[118]**^{I}	[59]^{T}, **[123]**^{T}, **[58]**$^{\dot{V}}$	**[118]**^{I}	[171], [149]^{T}, [155]^{I,T,V}, **[58]**$^{\dot{V}}$, **[124]**, [82]^{T}	[59]^{T}, **[123]**^{T}, **[148]**^{T}, **[126]**^{I}		
Greatest Hits [182]	**[143]**, **[139]**^{V}				[183]$^{\dot{T},V}$, [184]$^{\dot{T},V}$, [185]$^{\dot{T},V}$, **[132]**$^{\dot{V}}$, **[133]**$^{\dot{T},V}$	[174]$^{\dot{V}}$		
HJDB	**[50]**	**[50]**						
ImageHear [186]						[187]^{I}		
MACS [188]					[149]^{T}, [155]$^{I,\dot{T},V}$, [161]$^{\dot{T}}$			
Medley-solos-DB [189]					[155]^{I,T,V}			
MTG-Jamendo [190]			[58]$^{\dot{V}}$		[58]$^{\dot{V}}$			
Odeon Cinematic SFX	**[117]**^{T}		[59]^{T}		**[117]**^{T}	[59]^{T}		
Private	**[36]**, **[4]**, **[46]**, **[143]**, **[117]**^{T}, **[139]**^{V}, **[42]**, **[128]** **[95]**, **[120]**, **[50]**, **[140]**^{V}	**[56]**, **[50]**, [152]^{T}	[191]^{T}	**[64]**^{T}	**[77]**^{T}, **[117]**^{T}, **[78]**, [192], [193]^{T}, [152]^{T}	**[86]**, [191]^{T}	[104]^{T}, **[55]**, [194]	**[96]**, **[94]**
RWCP-SSD-Onomatopoeia						**[103]**^{V}, **[146]**^{T}	**[146]**^{T}	
Sonniss	**[117]**^{T}		[59]^{T}		[155]^{I,T,V}, **[117]**^{T}	[59]^{T}		
SoundBible		[151]$^{\dot{T}}$, [52]$^{\dot{T}}$			[151]$^{\dot{T}}$, [52]$^{\dot{T}}$, [161]$^{\dot{T}}$			
Soundideas	**[39]**							
TUT Acoustic Scenes [195]	**[44]**				[149]$^{\dot{T}}$, **[131]**			
Unspecified	[196], **[40]**, [197], **[119]**	**[49]**, **[142]**, **[93]**			[196], [198], [199]^{T}	**[135]**, [197]	**[136]**, [200]$^{T,\dot{V}}$, [201], [202]$^{\dot{T}}$, [194]	**[93]**
UrbanSound8K [203]	**[118]**^{I}, **[44]**	**[118]**^{I}	[60], **[123]**^{T}	**[118]**^{I}	[171], [149]^{T}, [155]$^{I,\dot{T},V}$, **[131]**, **[124]**, [161]$^{\dot{T}}$	[60], **[123]**^{T}, **[126]**^{I}		
VAS [204]	[205]$^{\dot{V}}$, [206]^{I}		[206]^{I}					
VEGAS [207]	**[137]**^{V}, **[139]**$^{\dot{V}}$				**[137]**$^{\dot{V}}$			
VGGSound [208]	[206]^{I}, [209], [66]^{T}, [210]	[151]^{T}, [210]	[206]^{I}, [59]^{T}, [209]	[66]^{T}, [210]	[29]$^{\dot{T},V}$, [167]$^{\dot{V}}$, [171], [211]$^{I,\dot{T},V}$, [151]$^{\dot{T}}$, [212]$^{\dot{T},\dot{V}}$, [168]$^{\dot{V}}$, [184]$^{\dot{T},V}$ [185]$^{T,\dot{V}}$, [156]^{T}, [169]$^{T,\dot{V}}$, [69]$^{\dot{V}}$, [157]$^{T,\dot{V}}$, [159]$^{T,\dot{V}}$, [75]$^{\dot{V}}$, [213]$^{T,\dot{V}}$ [74]^{T}, [66]^{T}, [163]$^{T,\dot{V}}$	[187]^{I}, [167]$^{\dot{V}}$, [59]^{T}, [61]$^{\dot{V}}$	[6]$^{\dot{V}}$, [61]$^{\dot{V}}$, **[116]**^{I}, [81]$^{T,\dot{V}}$	
VocalSketch [214]					**[77]**$^{\dot{T}}$			
WavCaps [215]		[65]$^{\dot{T}}$, [52]^{T}			[29]$^{\dot{T},V}$, [149]$^{\dot{T}}$, [65]^{T}, [52]$^{\dot{T}}$, [156]$^{\dot{T}}$, [163]$^{\dot{T},V}$		[105]^{T}	
WavText5Ks [216]					[149]$^{\dot{T}}$, [155]$^{I,\dot{T},V}$			
WeSoundEffects	**[117]**^{T}		[59]^{T}		[155]$^{I,\dot{T},V}$, **[117]**^{T}	[59]$^{\dot{T}}$		
Zapsplat	**[43]**	**[141]**, **[130]**						

TABLE III
SUMMARY OF EVALUATION METHODS AND WORKS THAT EMPLOY THEM.

Section	Metric	Works that use it
Objective (Direct difference)	MSE/MAE (STFT/Mel)	[41], [50], [56], [70], [77], [82], [93], [94], [115], [122], [128], [131], [148], [185]
	Multi-scale	[93], [96]
	MCD	[105]
	Spectral Flux	[94]
Objective (Statistical difference)	KLD [106]	[5], [6], [29], [53], [59], [62], [66], [67], [72], [74]–[76], [79], [80], [82], [91], [92], [132], [149], [154]–[156], [158], [161]–[163], [167], [169], [186], [187], [209], [210], [212], [213]
	FAD [107]	[5], [6], [29], [39], [41]–[43], [52], [53], [59], [61], [63], [64], [66], [67], [71]–[80], [82], [92], [93], [95], [96], [104], [105], [116]–[120], [122]–[131], [134], [141], [143], [149], [151], [152], [154], [156], [158], [159], [161]–[164], [167]–[169], [171], [176], [178], [183], [184], [186], [187], [191], [192], [209]–[213], [217], [218]
Neural evaluation	VGGish Loss [108]	[60], [79], [97], [196]
	CLAP [109]	[29], [53], [72], [79], [80], [92], [104], [149]–[151], [155], [156], [161], [178], [193], [199], [210], [211]
Subjective-equivalent	ViSQOL [110]	[56], [58], [105]
	PEAQ [112]	[56], [139]
Subjective	MOS / MUSHRA(-like) [113], [114]	[4], [36], [38], [39], [43]–[45], [55], [58], [59], [62], [66], [72], [73], [75], [77], [80], [86], [93], [103], [105], [115], [117], [122], [127], [136], [138], [143], [145], [146], [148], [149], [152], [155], [160], [165], [171], [174], [182], [185], [193], [202], [205], [210], [211]

the current literature suggests a more nuanced reality. The choice of generative paradigm often appears driven by broader trends (e.g., the favourability of Diffusion models for high-fidelity generation) and data availability, rather than a specific optimization for categories like "Lifeforms" or "Machinery". Nonetheless, Table I reveals the breadth of application for generative paradigms like GANs and Diffusion across diverse taxonomies, while also highlighting potential gaps where certain sound types may be relatively under-explored using the latest techniques. This lack of specialization suggests ample room for research into architecture-sound suitability.

Similarly, examining the interplay between datasets and generative paradigms (Table II) reveals important patterns. The prevalence of certain models (e.g., Diffusion) in conjunction with large benchmark datasets like AudioSet or DCASE subsets suggests these architectures scale effectively and are the focus of comparative academic study. Conversely, the significant number of studies relying on private or unspecified datasets, while potentially enabling highly specific task tuning, introduces challenges in reproducibility and comparative assessment. This "dataset divide" further complicates the creation of a universal cookbook, as performance is intrinsically tied to the training data's characteristics, which are not always transparent or accessible. The frequent use of private data, often marked in bold in Table II, underscores a critical challenge for standardized benchmarking in the field.

For all these reasons, navigating this scope review depends on the reader's profile. For practitioners (sound designers, game developers, producers) seeking to leverage Neural Synthesis of Sound Effects, the insights point towards exploring recent, high-performing generative paradigms, particularly those demonstrating strong subjective evaluations or utilizing relevant conditioning (e.g., text prompts for specific effects). Tables I and II can serve as a guide for identifying papers that tackle similar sound types or use potentially relevant (public) datasets as a starting point. For AI researchers, this review charts the evolving landscape of current research, illuminating the plethora of opportunities that emerge throughout the ongoing exploration rather than focusing solely on final outcomes. The tables reveal under-explored intersections of taxonomies and generative principles, the persistent challenge of controllable synthesis despite advancements in conditioning, the need for robust evaluation beyond standard metrics (especially for perceptual attributes crucial to sound effects), and the critical issue of dataset accessibility and standardization. Investigating why certain paradigms excel, developing more efficient models, and bridging the gap between benchmark performance and real-world usability remain key research avenues. Ultimately, while a simple recipe is absent, this synthesis provides the necessary ingredients and context for both applying and advancing the state-of-the-art in Neural Sound Effect Synthesis.

X. CONCLUSIONS

This literature scope review reveals a growing interest in generating sound effects through NAS. The rise of generative models has improved sound synthesis' plausibility, consistency, and diversity. These gains primarily concern synthesis workflows; broader 'sound creation' practices (recording, layering, editing) remain predominant, and adoption is still emerging. Advancements have the potential to be revolutionary in various industries, including film, theater, music, video games, and more. The generative potential that these models offer goes beyond the limitations of traditional, pre-recorded sound databases. It opens up a more comprehensive array of possibilities for artists, enabling them to sidestep the well-known issue of repetitive sound effects, such as the "machine-gun" effect, where the same sound is used repeatedly. Moreover, it enables the invention of novel sound effects by exploring combinations—e.g., hybrid material timbres, non-physical morphs, and cross-domain textures—that are impractical or impossible to realize acoustically. Unlike handcrafted pipelines based on layering/processing recordings or bespoke procedural patches, NAS offers a learned generative manifold and conditional controls (text, video, parameters) that accelerate ideation and yield candidates for human curation.

The current state of the field demands high precision and control for widespread application. This limitation has spurred a trend towards multimodal approaches, enabling more nuanced and refined sound synthesis. Although the sound produced by these methods may not be final, it serves as a crucial tool in the iterative process of sound design, supporting the synthesis of more polished and practical sound effects.

In conducting our review, we adhered to the PRISMA principles and harnessed the latest advancements in LLMs,

to significantly enhance our filtering process beyond what is achievable through standard keyword-based searches. This approach, though cutting-edge, comes with inherent challenges due to the lack of consensus for the standardisation of the process. It specifically affects the comprehensively covering all relevant literature. Despite our diligent efforts, some articles may have been overlooked. To mitigate this, we have created an online platform where readers can suggest any missing works for future inclusion. This collaborative effort will help ensure our review remains as comprehensive and up-to-date as possible[4].

XI. Acknowledgments

Activities described in this contribution were partially funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 101003750, the Ministry of Economy and Competitiveness of Spain under grant PID2021-128469OB-I00.

References

[1] S. H. Hawley, V. Chatziioannou, and A. Morrison, "Synthesis of musical instrument sounds: Physics-based modeling or machine learning?" *Phys. Today*, vol. 16, no. 1, pp. 20–28 (2020).

[2] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, *et al.*, "Neural audio synthesis of musical notes with wavenet autoencoders," presented at the *International Conference on Machine Learning*, pp. 1068–1077 (2017).

[3] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," presented at the *International Conference on Learning Representations* (2020).

[4] M. Chang, Y. R. Kim, and G. J. Kim, "A Perceptual Evaluation of Generative Adversarial Network Real-Time Synthesized Drum Sounds in a Virtual Environment," presented at the *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 144–148 (2018 Dec.).

[5] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, *et al.*, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," *Proceedings of the International Conference on Machine Learning*, pp. 21450–21474 (2023).

[6] X. Mei, V. Nagaraja, G. Le Lan, Z. Ni, E. Chang, Y. Shi, *et al.*, "Foleygen: Visually-guided audio generation," presented at the *IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6 (2024).

[7] K. P. Murphy, *Probabilistic machine learning: Advanced topics* (MIT press, 2023).

[8] A. Farnell, *Designing sound* (Mit Press, 2010).

[9] P. R. Cook, *Real sound synthesis for interactive applications* (CRC Press, 2002).

[10] T. Wilmering, D. Moffat, A. Milo, and M. B. Sandler, "A history of audio effects," *Applied Sciences*, vol. 10, no. 3, p. 791 (2020).

[11] BBC, "The BBC Year Book 1931, chapter 'The Use of Sound Effects'," *British Broadcasting Corporation*, pp. 194–197 (1931).

[12] A. Farnell, "An introduction to procedural audio and its application in computer games," presented at the *Audio mostly conference*, vol. 23, pp. 1–31 (2007).

[13] European Commission, Directorate-General for Research and Innovation, Directorate E-Prosperity, Unit E4 – Industry 5.0 & AI in Science, "Living guidelines on the responsible use of generative AI in research," (2025).

[14] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71 (2021).

[15] D. Rocchesso and F. Fontana, *The sounding object* (Mondo estremo, 2003).

[16] R. Viers, *The Sound Effects Bible: how to create and record Hollywood style sound effects* (Michael Wiese Productions, 2008).

[17] D. Moffat, R. Selfridge, and J. D. Reiss, "Sound effect synthesis," in *Foundations in Sound Design for Interactive Media*, pp. 274–299 (Routledge, 2019).

[18] D. Foster, *Generative deep learning* ("O'Reilly Media, Inc.", 2022).

[19] Y. Zhao, X. Xia, and R. Togneri, "Applications of Deep Learning to Audio Generation," *IEEE Circuits and Systems Magazine*, vol. 19, no. 4, pp. 19–38 (2019).

[20] A. Natsiou and S. O'Leary, "Audio representations for deep learning in sound synthesis: A review," presented at the *IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–8 (2021).

[21] M. Božić and M. Horvat, "A survey of deep learning audio generation methods," *arXiv preprint arXiv:2406.00146* (2024).

[22] D. Menexopoulos, P. Pestana, and J. Reiss, "The state of the art in procedural audio," *Journal of the Audio Engineering Society*, vol. 71, no. 12, pp. 826–848 (2023).

[23] B. Hayes, J. Shier, G. Fazekas, A. McPherson, and C. Saitis, "A review of differentiable digital signal processing for music and speech synthesis," *Frontiers in Signal Processing*, vol. 3 (2024).

[24] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," *arXiv preprint arXiv:2011.06801* (2020).

[25] A. Vinay and A. Lerch, "Evaluating generative audio systems and their metrics," presented at the *Proceedings of the 23rd International Society for Music Information Retrieval Conference* (2022).

[26] A. Barahona-Ríos, *Deep Learning for the Synthesis of Sound Effects*, Ph.D. thesis, University of York (2023).

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, *et al.*, "WaveNet: A Generative Model for Raw Audio," presented at the *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, p. 125 (2016).

[28] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, *et al.*, "SampleRNN: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837* (2016).

[29] Z. Chen, P. Seetharaman, B. Russell, O. Nieto, D. Bourgin, A. Owens, *et al.*, "Video-Guided Foley Sound Generation with Multimodal Controls," presented at the *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, p. 18770–18781 (2025 Jun.).

[30] L. Tunstall, L. Von Werra, and T. Wolf, *Natural language processing with transformers* ("O'Reilly Media, Inc.", 2022).

[31] F. Dennstädt, J. Zink, P. M. Putora, J. Hastings, and N. Cihoric, "Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain," *Systematic Reviews*, vol. 13, no. 1, p. 158 (2024).

[32] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research*, vol. 22, no. 57, pp. 1–64 (2021).

[33] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The Curious Case of Neural Text Degeneration," presented at the *International Conference on Learning Representations* (2020).

[34] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," presented at the *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410 (2019).

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144 (2020).

[36] A. Barahona-Rıos and S. Pauletto, "Synthesising knocking sound effects using conditional WaveGAN," presented at the *17th Sound and Music Computing Conference* (2020).

[37] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," presented at the *Proceedings of the AAAI conference on artificial intelligence*, vol. 32 (2018).

[38] A. Lavault, A. Roebel, and M. Voiry, "StyleWaveGAN: Style-based synthesis of drum sounds with extensive controls using generative adversarial networks," *arXiv preprint arXiv:2204.00907* (2022).

[39] Y. Liu and C. Jin, "Conditional sound effects generation with regularized wgan," presented at the *20th Sound and Music Computing Conference* (2023).

[40] J. Drysdale, M. Tomczak, and J. Hockman, "Style-based drum synthesis with GAN inversion," presented at the *22nd Int. Society for Music Information Retrieval Conference* (2021).

[41] Y. Liu and C. Jin, "ICGAN: An Implicit Conditioning Method for Interpretable Feature Control of Neural Audio Synthesis," (2024 Jun.).

[4]https://mateocamara.github.io/neural-sound-effects/

[42] J. Nistal, S. Lattner, and G. Richard, "Drumgan: Synthesis of drum sounds with timbral feature conditioning using generative adversarial networks," *arXiv preprint arXiv:2008.12073* (2020).

[43] M. Comunità, H. Phan, and J. D. Reiss, "Neural synthesis of footsteps sound effects with generative adversarial networks," *arXiv preprint arXiv:2110.09605* (2021).

[44] A. Madhu *et al.*, "EnvGAN: Adversarial Synthesis of Environmental Sounds for Data Augmentation," *arXiv preprint arXiv:2104.07326* (2021).

[45] A. Barahona-Ríos and T. Collins, "SpecSinGAN: Sound effect variation synthesis using single-image GANs," *arXiv preprint arXiv:2110.07311* (2021).

[46] J. Drysdale, M. Tomczak, and J. Hockman, "Adversarial synthesis of drum sounds," presented at the *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (2020).

[47] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114* (2013).

[48] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," presented at the *International conference on machine learning*, pp. 1530–1538 (2015).

[49] C. Aouameur, P. Esling, and G. Hadjeres, "Neural drum machine: An interactive system for real-time synthesis of drum sounds," *arXiv preprint arXiv:1907.02637* (2019).

[50] M. Tomczak, M. Goto, and J. Hockman, "Drum Synthesis and Rhythmic Transformation with Adversarial Autoencoders," presented at the *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2427–2435 (2020 Oct.).

[51] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023).

[52] T. Karchkhadze, H. S. Kavaki, M. R. Izadi, B. Irvin, M. Kegler, A. Hertz, *et al.*, "Latent clap loss for better foley sound synthesis," presented at the *32nd European Signal Processing Conference (EUSIPCO)*, pp. 351–355 (2024).

[53] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," presented at the *Forty-first International Conference on Machine Learning* (2024).

[54] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011* (2021).

[55] A. Ramires, P. Chandna, X. Favory, E. Gómez, and X. Serra, "Neural percussive synthesis parameterised by high-level timbral features," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 786–790 (2020).

[56] M. Cámara, J. L. Blanco Murillo, and D. Project, "Phase-Aware Transformations in Variational Autoencoders for Audio Effects," *Journal of the Audio Engineering Society*, pp. 731–741 (2022 Sep.).

[57] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30 (2017).

[58] R. San Roman, Y. Adi, A. Deleforge, R. Serizel, G. Synnaeve, and A. Defossez, "From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion," *Advances in Neural Information Processing Systems*, vol. 36, pp. 1526–1538 (2023 Dec.).

[59] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, *et al.*, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352* (2022).

[60] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," presented at the *IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6 (2021).

[61] S. Pascual, C. Yeh, I. Tsiamas, and J. Serrà, "Masked generative video-to-audio transformers with enhanced synchronicity," presented at the *European Conference on Computer Vision*, pp. 247–264 (2024).

[62] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, *et al.*, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733 (2023).

[63] M. Comunità, Z. Zhong, A. Takahashi, S. Yang, M. Zhao, K. Saito, *et al.*, "Specmaskgit: Masked generative modeling of audio spectrograms for efficient audio synthesis and beyond," *arXiv preprint arXiv:2406.17672* (2024).

[64] S. Andreu and M. Villanueva Aylagas, "Neural Synthesis of Sound Effects Using Flow-Based Deep Generative Models," *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 18, no. 1, pp. 2–9 (2022 Oct.).

[65] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. A. Bagherzadeh, C. Li, *et al.*, "Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," *arXiv preprint arXiv:2412.21037* (2024).

[66] Y. Wang, W. Guo, R. Huang, J. Huang, Z. Wang, F. You, *et al.*, "Frieren: Efficient Video-to-Audio Generation Network with Rectified Flow Matching," presented at the *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024).

[67] W. Guan, K. Wang, W. Zhou, Y. Wang, F. Deng, H. Wang, *et al.*, "Lafma: A latent flow matching model for text-to-audio generation," *InterSpeech* (2024).

[68] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851 (2020).

[69] S. Luo, C. Yan, C. Hu, and H. Zhao, "Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 48855–48876 (2023).

[70] Z. Guo, J. Mao, R. Tao, L. Yan, K. Ouchi, H. Liu, *et al.*, "Audio Generation with Multiple Conditional Diffusion Model," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 18153–18161 (2024 Mar.).

[71] G. Zhu, Y. Wen, M.-A. Carbonneau, and Z. Duan, "EDMSound: Spectrogram Based Diffusion Models for Efficient and High-Quality Audio Synthesis," *arXiv preprint arXiv:2311.08667* (2023).

[72] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, *et al.*, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *ACM Transactions on Audio, Speech, and Language Processing* (2024).

[73] A. Qi, X. Xie, and J. Wang, "Mtdiffusion: Multi-Task Diffusion Model With Dual-Unet for Foley Sound Generation," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 461–465 (2024 Apr.).

[74] I. Viertola, V. Iashin, and E. Rahtu, "Temporally Aligned Audio for Video with Autoregression," presented at the *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025 4).

[75] T. X. Pham, T. Ton, and C. D. Yoo, "MDSGen: Fast and Efficient Masked Diffusion Temporal-Aware Transformers for Open-Domain Sound Generation," presented at the *Intl. Conf. Learning Representations (ICLR)* (2025).

[76] J. Michaels, J. B. Li, L. Yao, L. Yu, Z. Wood-Doughty, and F. Metze, "Audio-Journey: Open Domain Latent Diffusion Based Text-To-Audio Generation," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6960–6964 (2024 Apr.).

[77] Y. Chung, J. Lee, and J. Nam, "T-Foley: A Controllable Waveform-Domain Diffusion Model for Temporal-Event-Guided Foley Sound Synthesis," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6820–6824 (2024).

[78] S. Rouard and G. Hadjeres, "CRASH: Raw audio score-based generative modeling for controllable high-resolution drum sound synthesis," *arXiv preprint arXiv:2106.07431* (2021).

[79] K. Saito, D. Kim, T. Shibuya, C.-H. Lai, Z. Zhong, Y. Takida, *et al.*, "Soundctm: Uniting score-based and consistency models for text-to-sound generation," presented at the *Audio Imagination: NeurIPS Workshop AI-Driven Speech, Music, and Sound Generation* (2024).

[80] H. Liu, R. Huang, Y. Liu, H. Cao, J. Wang, X. Cheng, *et al.*, "Audiolcm: Text-to-audio generation with latent consistency models," *arXiv preprint arXiv:2406.00356* (2024).

[81] Y. Zhang, Y. Gu, Y. Zeng, Z. Xing, Y. Wang, Z. Wu, *et al.*, "Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds," *arXiv preprint arXiv:2407.01494* (2024).

[82] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian, *et al.*, "AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 71340–71357 (2023 Dec.).

[83] Q. Kong, Y. Xu, T. Iqbal, Y. Cao, W. Wang, and M. D. Plumbley, "Acoustic Scene Generation with Conditional Samplernn," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 925–929 (2019 May).

[84] A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45 (2012).

[85] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," presented at the *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical* (2014).

[86] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, *et al.*, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," (2017 Feb.).

[87] R. Valle, R. Badlani, Z. Kong, S.-g. Lee, A. Goel, S. Kim, *et al.*, "Fugatto 1: Foundational generative audio transformer opus 1," presented at the *The Thirteenth International Conference on Learning Representations* (2025).

[88] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438* (2022).

[89] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460 (2020).

[90] Y. K. Singla, J. Shah, C. Chen, and R. R. Shah, "What do audio transformers hear? probing their representations for language delivery & structure," presented at the *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 910–925 (2022).

[91] X. Niu, J. Zhang, C. Walder, and C. P. Martin, "SoundLoCD: An Efficient Conditional Discrete Contrastive Latent Diffusion Model for Text-to-Sound Generation," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–265 (2024).

[92] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Retrieval-Augmented Text-to-Audio Generation," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 581–585 (2024 Apr.).

[93] Y. Liu, C. Jin, and D. Gunawan, "DDSP-SFX: Acoustically-guided sound effects generation with differentiable digital signal processing," *arXiv preprint arXiv:2309.08060* (2023).

[94] J. Shier, F. Caspe, A. Robertson, M. Sandler, C. Saitis, and A. McPherson, "Differentiable modelling of percussive audio with transient and spectral synthesis," *arXiv preprint arXiv:2309.06649* (2023).

[95] A. Ramires, J. Juras, J. D. Parker, and X. Serra, "A study of control methods for percussive sound synthesis based on gans," presented at the *Proceedings of the 25th International Conference on Digital Audio Effects (DAFx20in22)*, pp. 224–231 (2022).

[96] A. Barahona-Ríos and T. Collins, "NoiseBandNet: controllable time-varying neural synthesis of sound effects using filterbanks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1573–1585 (2024).

[97] M. Hagiwara, M. Cusimano, and J.-Y. Liu, "Modeling animal vocalizations through synthesizers," *arXiv preprint arXiv:2210.10857* (2022).

[98] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017).

[99] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," presented at the *Proceedings of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 119–132 (2019).

[100] F. Font, G. Roma, and X. Serra, "Freesound technical demo," presented at the *Proceedings of the 21st ACM international conference on Multimedia*, pp. 411–412 (2013).

[101] BBC, "BBC Sound Effects Archive," *sound-effects.bbcrewind* (2024).

[102] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, Y. Yamashita, *et al.*, "Onoma-to-wave: Environmental sound synthesis from onomatopoeic words," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1 (2022).

[103] H. Ohnaka, S. Takamichi, K. Imoto, Y. Okamoto, K. Fujii, and H. Saruwatari, "Visual Onoma-to-Wave: Environmental Sound Synthesis from Visual Onomatopoeias and Sound-Source Images," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023 Jun.).

[104] H. F. García, O. Nieto, J. Salamon, B. Pardo, and P. Seetharaman, "Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025).

[105] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, *et al.*, "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704* (2023).

[106] M. Thomas and A. T. Joy, *Elements of information theory* (Wiley-Interscience, 2006).

[107] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Frèchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466* (2018).

[108] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, *et al.*, "CNN architectures for large-scale audio classification," presented at the *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 131–135 (2017).

[109] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023).

[110] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," presented at the *twelfth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6 (2020).

[111] M. Cámara, J. L. Blanco, and J. D. Reiss, "Parameter optimisation for a physical model of the vocal system," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 1, p. 27 (2025).

[112] M. Salovarda, I. Bolkovac, and H. Domitrovic, "Estimating perceptual audio system quality using PEAQ algorithm," presented at the *18th International Conference on Applied Electromagnetics and Communications*, pp. 1–4 (2005).

[113] International Telecommunication Union, "ITU-T Recommendation P.800: Methods for subjective determination of transmission quality," Technical report, International Telecommunication Union, Geneva, Switzerland (1996).

[114] International Telecommunication Union, "ITU-R Recommendation BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," Technical Report BS.1534-3, International Telecommunication Union, Geneva, Switzerland (2015 October).

[115] S. Ghose and J. J. Prevost, "AutoFoley: Artificial Synthesis of Synchronized Sound Tracks for Silent Videos with Deep Learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 1895–1907 (2021).

[116] H. Wang, J. Ma, S. Pascual, R. Cartwright, and W. Cai, "V2A-Mapper: A Lightweight Solution for Vision-to-Audio Generation by Connecting Foundation Models," presented at the *Proceedings of the AAAI Conference on Artificial Intelligence* (2024).

[117] M. Kang, S. Oh, H. Moon, K. Lee, and B. S. Chon, "FALL-E: A Foley Sound Synthesis Model and Strategies," *arXiv preprint arXiv:2306.09807* (2023).

[118] J. Lee, H. Nam, and Y.-H. Park, "VIFS: An end-to-end variational inference for foley sound synthesis," *arXiv preprint arXiv:2306.05004* (2023).

[119] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Conditional Foley Sound Synthesis with Limited Data: Two-Stage Data Augmentation Approach with StyleGAN2-ADA," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746 (2015 Oct.).

[120] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "DCASE Task-7: StyleGAN2-Based Foley Sound Synthesis," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746 (2015 Oct.).

[121] R. Scheibler, T. Hasumi, Y. Fujita, T. Komatsu, R. Yamamoto, and K. Tachibana, "Foley sound synthesis with a class-conditioned latent diffusion model," presented at the *DCASE-Workshop on Detection and Classification of Acoustic Scenes and Events* (2023).

[122] Z. Xie, B. Li, X. Xu, M. Wu, and K. Yu, "Enhancing Audio Generation Diversity with Visual Information," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 866–870 (2024).

[123] A. Pillay, S. Betko, A. Liloia, H. Chen, and A. Shah, "Exploring Domain-Specific Enhancements for a Neural Foley Synthesizer," *arXiv preprint arXiv:2309.04641* (2023).

[124] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Audio Diffusion for Foley Sound Synthesis," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746 (2015 Oct.).

[125] H. Zhang, K. Qian, L. Shen, L. Li, K. Xu, and B. Hu, "From noise to sound: Audio synthesis via diffusion models," *signal*, vol. 15, p. 16 (2023).

[126] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "High-Quality Foley Sound Synthesis Using Monte Carlo Dropout," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746 (2015 Oct.).

[127] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208* (2018).

[128] J. Nistal, C. Aouameur, I. Velarde, and S. Lattner, "DrumGAN VST: A plugin for drum sound analysis/synthesis with autoencoding generative adversarial networks," *arXiv preprint arXiv:2206.14723* (2022).

[129] Y. Liu and C. Jin, "ICGAN: An implicit conditioning method for interpretable feature control of neural audio synthesis," *arXiv preprint arXiv:2406.07131* (2024).

[130] Z.-S. Yang and J. Hockman, "A Plugin for Neural Audio Synthesis of Impact Sound Effects," presented at the *Proceedings of the 18th International Audio Mostly Conference*, pp. 143–146 (2023 Aug.).

[131] S. Pascual, G. Bhattacharya, C. Yeh, J. Pons, and J. Serrà, "Full-band general audio synthesis with score-based diffusion," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023).

[132] K. Su, K. Qian, E. Shlizerman, A. Torralba, and C. Gan, "Physics-Driven Diffusion Models for Impact Sound Synthesis From Videos," presented at the *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9749–9759 (2023 Jan.).

[133] Z. Xie, S. Yu, Q. He, and M. Li, "Sonicvisionlm: Playing sound with vision language models," presented at the *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26866–26875 (2024).

[134] Y. Yuan, H. Liu, X. Liu, X. Kang, M. D. Plumbley, and W. Wang, "Latent diffusion model based foley sound generation system for dcase challenge 2023 task 7," *arXiv preprint arXiv:2305.15905* (2023).

[135] H. Caracalla and A. Roebel, "Sound texture synthesis using convolutional neural networks," presented at the *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (2019).

[136] H. Caracalla and A. Roebel, "Sound texture synthesis using RI spectrograms," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 416–420 (2020).

[137] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan, "Generating Visually Aligned Sound from Videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302 (2020).

[138] S. Ghose and J. J. Prevost, "FoleyGAN: Visually Guided Generative Adversarial Network-Based Synchronous Sound Generation in Silent Videos," (2021 Jul.).

[139] S. Liu, S. Li, and H. Cheng, "Towards an End-to-End Visual-to-Raw-Audio Generation With GAN," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1299–1312 (2022 Mar.).

[140] K. Uchiyama and K. Kawamoto, "AudioVisual Model for Generating Eating Sounds Using Food ASMR Videos," *IEEE Access*, vol. 9, pp. 50106–50111 (2021).

[141] D. Serrano and M. Cartwright, "A General Framework for Learning Procedural Audio Models of Environmental Sounds," *arXiv preprint arXiv:2303.02396* (2023).

[142] M. Cámara and J. L. Blanco, "FOLEY-VAE: Generación de efectos de audio para cine con inteligencia artificial," *arXiv preprint arXiv:2310.15663* (2023).

[143] P. Kamath, C. Gupta, L. Wyse, and S. Nanayakkara, "Example-Based Framework for Perceptually Guided Audio Texture Generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2555–2565 (2024 Jan.).

[144] G. Li, X. Xu, L. Dai, M. Wu, and K. Yu, "Diverse and Vivid Sound Generation from Text Descriptions," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023 Jun.).

[145] Y. Okamoto, K. Imoto, S. Takamichi, R. Nagase, T. Fukumori, and Y. Yamashita, "Environmental sound synthesis from vocal imitations and sound event labels," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 411–415 (2024).

[146] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "Onomatowave: Environmental Sound Synthesis from Onomatopoeic Words," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1 (2022).

[147] S. Pauletto, A. Barahona-Rios, V. Madaghiele, and Y. Seznec, "Sonifying Energy Consumption Using SpecSinGAN," presented at the *Sound and Music Computing Conference* (2023 Jun.).

[148] V. M. O. Ramos and S. Lee, "Synthesis of Disparate Audio Species via Recurrent Neural Embedding," presented at the *IEEE International Symposium on Multimedia (ISM)*, pp. 198–201 (2023 Dec.).

[149] J. Huang, Y. Ren, R. Huang, D. Yang, Z. Ye, C. Zhang, *et al.*, "Make-an-audio 2: Temporal-enhanced text-to-audio generation," *arXiv preprint arXiv:2305.18474* (2023).

[150] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," presented at the *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572 (2024).

[151] M. Haji-Ali, W. Menapace, A. Siarohin, G. Balakrishnan, S. Tulyakov, and V. Ordonez, "Taming data and transformers for audio generation," *arXiv preprint arXiv:2406.19388* (2024).

[152] Z. Xie, X. Xu, Z. Wu, and M. Wu, "Picoaudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation," *arXiv preprint arXiv:2407.02869* (2024).

[153] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," presented at the *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025).

[154] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *arXiv preprint arXiv:2304.13731* (2023).

[155] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, *et al.*, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," presented at the *International Conference on Machine Learning*, pp. 13916–13932 (2023).

[156] C. Li, M. Xu, and D. Yu, "SRC-gAudio: Sampling-Rate-Controlled Audio Generation," presented at the *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–6 (2024).

[157] S. Mo, J. Shi, and Y. Tian, "Text-to-audio generation synchronized with videos," *arXiv preprint arXiv:2403.07938* (2024).

[158] S. Mo, J. Shi, and Y. Tian, "DiffAVA: Personalized text-to-audio generation with visual alignment," *arXiv preprint arXiv:2305.12903* (2023).

[159] S. Mo, J. Shi, and Y. Tian, "Text-to-audio generation synchronized with videos," *arXiv preprint arXiv:2403.07938* (2024).

[160] Y. Wang, H. Chen, D. Yang, Z. Wu, and X. Wu, "AudioComposer: Towards Fine-grained Audio Generation with Natural Language Descriptions," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025).

[161] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[162] M. Yang, B. Shi, M. Le, W.-N. Hsu, and A. Tjandra, "Audiobox TTA-RAG: Improving zero-shot and few-shot text-to-audio with retrieval-augmented generation," *arXiv preprint arXiv:2411.05141* (2024).

[163] Q. Yang, B. Mao, Z. Wang, X. Nie, P. Gao, Y. Guo, *et al.*, "Draw an audio: Leveraging multi-instruction for video-to-audio synthesis," *arXiv preprint arXiv:2409.06135* (2024).

[164] Y. Yuan, H. Liu, X. Liu, X. Kang, P. Wu, M. D. Plumbley, *et al.*, "Text-driven foley sound generation with latent diffusion model," *arXiv preprint arXiv:2306.10359* (2023).

[165] X. Liu, Z. Zhu, H. Liu, Y. Yuan, M. Cui, Q. Huang, *et al.*, "Wavjourney: Compositional audio creation with large language models," *arXiv preprint arXiv:2307.14335* (2023).

[166] A. McDonagh, J. Lemley, R. Cassidy, and P. Corcoran, "Synthesizing Game Audio Using Deep Neural Networks," presented at the *IEEE Games, Entertainment, Media Conference (GEM)*, pp. 1–9 (2018 Aug.).

[167] X. Cheng, X. Wang, Y. Wu, Y. Wang, and R. Song, "Lova: Long-form video-to-audio generation," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025).

[168] Z. Huang, D. Luo, J. Wang, H. Liao, Z. Li, and Z. Wu, "Rhythmic foley: A framework for seamless audio-visual alignment in video-to-audio synthesis," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025).

[169] B. Li, F. Yang, Y. Mao, Q. Ye, H. Chen, and Y. Zhong, "Tri-ergon: Fine-grained video-to-audio generation with multi-modal conditions and lufs control," presented at the *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 4616–4624 (2025).

[170] L. Zhang, S. Mo, Y. Zhang, and P. Morgado, "Audio-Synchronized Visual Animation," *arXiv preprint arXiv:2403.05659* (2024).

[171] M. F. Colombo, F. Ronchini, L. Comanducci, and F. Antonacci, "Mambafoley: Foley sound generation using selective state-space models," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025).

[172] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740 (2020).

[173] Y. Zhang, L. Shao, and C. G. Snoek, "Repetitive activity counting by sight and sound," presented at the *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14070–14079 (2021).

[174] Y. Du, Z. Chen, J. Salamon, B. Russell, and A. Owens, "Conditional generation of audio from video via foley analogies," presented at the *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2436 (2023).

[175] K. Choi, J. Im, L. Heller, B. McFee, K. Imoto, Y. Okamoto, *et al.*, "Foley sound synthesis at the dcase 2023 challenge," *arXiv preprint arXiv:2304.12521* (2023).

[176] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Foley Sound Synthesis Based on GAN using Contrastive Learning Without Label Information," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746 (2015 Oct.).

[177] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," presented at the *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012 (2022).

[178] C. Chen, P. Peng, A. Baid, Z. Xue, W.-N. Hsu, D. Harwath, *et al.*, "Action2sound: Ambient-aware generation of action sounds from egocentric videos," presented at the *European Conference on Computer Vision*, pp. 277–295 (2024).

[179] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, *et al.*, "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100," *International Journal of Computer Vision (IJCV)*, vol. 130, p. 33–55 (2022).

[180] K. J. Piczak, "ESC: Dataset for environmental sound classification," presented at the *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018 (2015).

[181] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852 (2021).

[182] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," presented at the *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2405–2413 (2016).

[183] M. Comunità, R. F. Gramaccioni, E. Postolache, E. Rodolà, D. Comminiello, and J. D. Reiss, "Syncfusion: Multimodal Onset-Synchronized Video-to-Audio Foley Synthesis," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 936–940 (2024).

[184] M. Ishii, A. Hayakawa, T. Shibuya, and Y. Mitsufuji, "A Simple but Strong Baseline for Sounding Video Generation: Effective Adaptation of Audio and Video Diffusion Models for Joint Generation," *arXiv preprint arXiv:2409.17550* (2024).

[185] Y. Jeong, Y. Kim, S. Chun, and J. Lee, "Read, watch and scream! sound generation from text and video," presented at the *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 17590–17598 (2025).

[186] R. Sheffer and Y. Adi, "I hear your true colors: Image guided audio generation," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023).

[187] Y. Ahn, C. Wang, Y. Wu, J. W. Shin, and S. Liu, "GRAVO: Learning to Generate Relevant Audio from Visual Features with Noisy Online Videos," presented at the *Interspeech*, pp. 2743–2747 (2023).

[188] I. Martin Morato and A. Mesaros, "Diversity and bias in audio captioning datasets," *DCASE* (2021).

[189] V. Lostanlen, C.-E. Cella, R. Bittner, and S. Essid, "Medley-solos-DB: a cross-collection dataset for musical instrument recognition," *Zenodo* (2018).

[190] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo Dataset for Automatic Music Tagging," presented at the *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)* (2019).

[191] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "The X-Lance System for DCASE2023 Challenge Task 7: Foley Sound Synthesis Track B," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746 (2015 Oct.).

[192] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Foley Sound Synthesis in Waveform Domain with Diffusion Model," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746 (2015 Oct.).

[193] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, *et al.*, "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821* (2023).

[194] R. Takizawa and S. Hirai, "Synthesis of Explosion Sounds from Utterance Voice of Onomatopoeia Using Transformer," presented at the *28th International Conference on Intelligent User Interfaces*, pp. 87–90 (2023 Mar.).

[195] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," presented at the *24th European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132 (2016).

[196] J. M. Antognini, M. Hoffman, and R. J. Weiss, "Audio Texture Synthesis with Random Neural Networks: Improving Diversity and Quality," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3587–3591 (2019 May).

[197] W. Liu, G. Liu, X. Ji, J. Zhai, and Y. Dai, "Sound Texture Generative Model Guided by a Lossless Mel-Frequency Convolutional Neural Network," *IEEE Access*, vol. 6, pp. 48030–48041 (2018).

[198] J. Antognini, M. Hoffman, and R. J. Weiss, "Synthesizing diverse, high-quality audio textures," *arXiv preprint arXiv:1806.08002* (2018).

[199] Z. Chen, D. Geng, and A. Owens, "Images that sound: Composing images and sounds on a single canvas," *Advances in Neural Information Processing Systems*, vol. 37, pp. 85045–85073 (2024).

[200] G. Chen, G. Wang, X. Huang, and J. Sang, "Semantically consistent video-to-audio generation using multimodal language large model," *arXiv preprint arXiv:2404.16305* (2024).

[201] X. Jin, S. Li, T. Qu, D. Manocha, and G. Wang, "Deep-modal: real-time impact sound synthesis for arbitrary shapes," presented at the *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1171–1179 (2020).

[202] X. Su, J. E. Froehlich, E. Koh, and C. Xiao, "SonifyAR: Context-Aware Sound Generation in Augmented Reality," (2024 May).

[203] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," presented at the *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044 (2014).

[204] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan, "Generating visually aligned sound from videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302 (2020).

[205] C. Cui, Z. Zhao, Y. Ren, J. Liu, R. Huang, F. Chen, *et al.*, "VarietySound: Timbre-Controllable Video to Sound Generation Via Unsupervised Information Disentanglement," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023 Jun.).

[206] V. Iashin and E. Rahtu, "Taming visually guided sound generation," *arXiv preprint arXiv:2110.08791* (2021).

[207] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," presented at the *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3550–3558 (2018).

[208] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A Large-scale Audio-Visual Dataset," presented at the *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2020).

[209] X. Liu, K. Su, and E. Shlizerman, "Tell What You Hear From What You See–Video to Audio Generation Through Text," *arXiv preprint arXiv:2411.05679* (2024).

[210] M. Xu, C. Li, X. Tu, Y. Ren, R. Chen, Y. Gu, *et al.*, "Video-to-audio generation with hidden alignment," *arXiv preprint arXiv:2407.07464* (2024).

[211] H.-W. Dong, X. Liu, J. Pons, G. Bhattacharya, S. Pascual, J. Serrà, *et al.*, "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models," presented at the *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5 (2023 Oct.).

[212] Y. Hu, Y. Gu, C. Li, R. Chen, and D. Yu, "Video-to-Audio Generation with Fine-grained Temporal Semantics," *arXiv preprint arXiv:2409.14709* (2024).

[213] Y. Ren, C. Li, M. Xu, W. Liang, Y. Gu, R. Chen, *et al.*, "STA-V2A: Video-to-audio generation with semantic and temporal alignment," presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025).

[214] M. Cartwright and B. Pardo, "Vocalsketch: Vocally imitating audio concepts," presented at the *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 43–46 (2015).

[215] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, *et al.*, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[216] S. Deshmukh, B. Elizalde, and H. Wang, "Audio Retrieval with WavText5K and CLAP Training," *arXiv preprint arXiv:2209.14275* (2022).

[217] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746 (2015 Oct.).

[218] G. Kim, A. Martinez, Y.-C. Su, B. Jou, J. Lezama, A. Gupta, *et al.*, "A Versatile Diffusion Transformer with Mixture of Noise Levels for Audiovisual Generation," *arXiv preprint arXiv:2405.13762* (2024).