# PHYAUDIOBENCH: BENCHMARKING PHYSICAL CONSISTENCY IN AUDIO GENERATION THROUGH CONTROLLED INTERVENTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent text-to-audio generation models have achieved substantial progress in perceptual quality and semantic alignment, yet existing evaluation largely focuses on whether generated audio sounds plausible, matches the prompt, and follows the requested temporal structure. Whether generated sounds behave consistently with their described physical conditions remains much less characterized. We introduce **PhyAudioBench**, a hierarchical benchmark for evaluating physical consistency in general audio generation. PhyAudioBench evaluates acoustic and prompt fidelity, temporal structure, and physical consistency across 240 structured prompts spanning eight acoustic scenes, three complexity levels, seven physical properties, and twelve controlled intervention types. Rather than judging physical realism from isolated waveforms, we construct minimally modified prompt pairs that alter a target physical condition and test whether the corresponding acoustic evidence changes in the expected direction. A shared multimodal observer grounds acoustic events in time, while property-specific signal evaluators perform physical measurement. Evaluating ten representative audio generators with fifteen metrics reveals that physical-consistency scores are consistently lower than lower-level scores, all ten models change rank between Level 1 and Level 3, and opposite-direction responses constitute a prominent failure mode under controlled interventions. These results expose a gap between generating audio that sounds plausible and generating audio that behaves consistently in physical world.

## 1 INTRODUCTION

Recent advances in generative modeling have substantially improved the fidelity, diversity, and controllability of audio generation (Liu et al., 2024; Evans et al., 2024; Hung et al., 2024). Evaluation has progressed accordingly. Speech and music generation benefit from specialized linguistic, paralinguistic, perceptual, and structural measures (Yang et al., 2021; Huang et al., 2023a; Yuan et al., 2023; Melechovsky et al., 2024), while recent benchmarks for *general audio generation* increasingly evaluate fine-grained instruction following, including sound-event identity, event count, temporal ordering, and timestamp consistency (Liu et al., 2026; Tian et al., 2026). These efforts provide increasingly strong measurements of *what* sounds are generated and *when* they occur. Yet they leave another important dimension insufficiently examined: whether the generated acoustic behavior is physically consistent with the described world.

General audio is often an acoustic observation of events unfolding under particular physical conditions. Its realization depends not only on the sound source itself, but also on source motion and distance, the surrounding environment, and propagation conditions. A model may therefore generate the correct event at the correct time while producing acoustic behavior that contradicts the prompt: changes in distance, motion, or environment should induce corresponding changes in attenuation, frequency, reverberation, or other acoustic cues. Such outputs can remain perceptually plausible and semantically correct while still being physically inconsistent.

Physical consistency is also substantially harder to evaluate than semantic or temporal correctness. The expected acoustic response is context dependent, multiple physical effects are entangled in the same waveform, and different phenomena manifest through different acoustic cues. A single generic



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

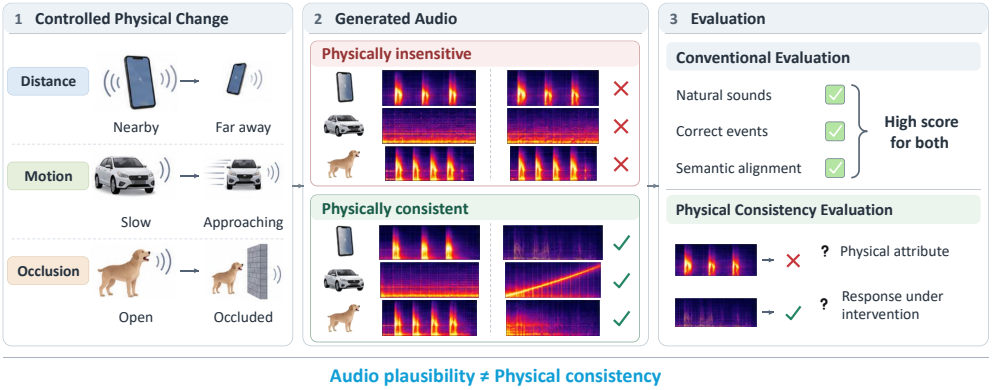



Figure 1: Semantic and temporal correctness do not necessarily imply physical consistency. Conventional evaluation asks whether the intended event is generated and when it occurs, while physical evaluation further asks whether its acoustic behavior responds consistently to the described physical conditions.

physical-realism score is therefore difficult to interpret. Although multimodal audio models provide flexible open-ended audio understanding (Sun et al., 2026; Wang et al., 2026), directly using a learned judge for physical realism further entangles the evaluation with the judge's own perception, reasoning, and calibration.

We therefore move beyond *absolute judgment* by evaluating *controlled response*. In addition to asking whether a generated waveform is physically realistic in isolation, we ask whether its measurable acoustic behavior changes appropriately when one physical condition is changed while the remaining semantic content is preserved. Given a prompt $p$, we construct a minimally modified counterpart $p'$ that intervenes on a target physical factor. For example, changing a source from "nearby" to "far away" should produce a corresponding change in distance-related acoustic evidence. This directly probes whether a generator is responsive to the physical condition that should govern its output.

Building on this idea, we introduce **PhyAudioBench**, a hierarchical benchmark for evaluating general audio generation at progressively stronger levels of consistency. **Level 1: Acoustic and Prompt Fidelity** evaluates whether the intended acoustic content is generated; **Level 2: Temporal Structure** evaluates event completeness and temporal organization; and **Level 3: Physical Consistency** evaluates whether localized events exhibit acoustic behavior compatible with their physical conditions. At Level 3, *attribute realization* measures physical acoustic evidence within an individual generation, while *controlled response* evaluates how that evidence changes under controlled interventions.

PhyAudioBench represents each instance with a structured acoustic specification that associates events with temporal relations, physical properties, and evaluation targets. We develop dedicated evaluators for seven physical properties (*distance, Doppler effect, reverberation, decay, resonance, echo, and occlusion*) and separate *perception* from *measurement*: a multimodal audio model identifies what events are audible and when they occur, while final scores are computed by deterministic metrics or property-specific acoustic evaluators. The benchmark contains 240 structured prompts across 8 acoustic scenes, covering 7 physical properties and 12 controlled intervention types.

Evaluating 10 representative text-to-audio models with 15 metrics reveals a substantial gap between acoustic plausibility and physical consistency. Across all models, scores consistently follow L1 $>$ L2 $>$ L3. Moreover, all ten models change rank between Level 1 and Level 3, with a Spearman rank correlation of only 0.370. Controlled-intervention analysis further shows that opposite-direction changes are more common than weak but correctly directed responses among measurable intervention failures. Together, these results expose a clear gap between generating audio that *sounds right* and generating audio that *behaves right*.

## 2 From Acoustic Plausibility to Physical Consistency

**Physical consistency beyond acoustic plausibility.** General audio is not only a collection of recognizable sound events, but also an acoustic observation of events occurring under particular phys-

ical conditions. Even when the event is fixed, its acoustic realization varies with source motion, source–listener distance, sound propagation, surrounding environments, and interactions with materials or obstacles. These conditions are reflected through observable cues such as attenuation, spectral variation, reverberation, decay, resonance, or delayed reflections.

Consequently, semantic and temporal correctness do not guarantee physical consistency. A generated sample may contain the correct event at the correct time while exhibiting acoustic behavior incompatible with the described world. As illustrated in Fig. 1, two outputs can therefore appear equally plausible under conventional evaluation yet differ fundamentally in their physical behaviors. We use *physical consistency* to denote whether the observable acoustic behavior of a generated event is compatible with the physical conditions under which that event is specified to occur.

### 2.1 Controlled Response Beyond Absolute Judgement

Physical consistency cannot be fully assessed from an isolated waveform. The expected strength of an acoustic cue is highly context dependent, while different physical causes are entangled in the same signal and may produce similar spectral or temporal effects. Moreover, physical phenomena such as distance, Doppler effect, reverberation, echo, and occlusion manifest through different acoustic evidence. These properties make both a universal physical-realism metric and an unrestricted learned-judge score difficult to interpret reliably.

We therefore complement *absolute judgement* with *controlled response*. Given a prompt $p$, we construct a minimally modified counterpart $p'$ that changes only one target physical condition while preserving the sound source, event structure, and other semantic content as much as possible. Let $x$ and $x'$ denote the corresponding generations, and let $f_a$ measure acoustic evidence for a physical attribute $a$. We evaluate the observed response

$$\Delta_a = f_a(x') - f_a(x). \tag{1}$$

For an intervention-affected attribute, the intervention specifies an expected direction $\delta_a \in \{-1, +1\}$, indicating whether the corresponding acoustic evidence should decrease or increase. Attributes intended to remain invariant are treated separately through static attribute realization rather than the controlled-response score. This controlled comparison tests whether the generator is actually responsive to the physical condition, rather than merely producing a plausible acoustic cue by chance or as a learned prior. It also avoids requiring a unique ground-truth waveform, as the key question is whether the measurable response follows the expected physical direction.

### 2.2 A Layered View of Evaluation

Physical evaluation naturally depends on lower-level acoustic evidence. Before measuring how an event behaves physically, the event must first be generated successfully and localized in time. This induces a layered evaluation process:

$$\underbrace{\text{what is generated}}_{\text{acoustic / prompt fidelity}} \rightarrow \underbrace{\text{when it occurs}}_{\text{temporal grounding}} \rightarrow \underbrace{\text{how it behaves}}_{\text{physical consistency}}. \tag{2}$$

Accordingly, semantic perception and temporal localization provide the event-level grounding required for subsequent physical measurement, while heterogeneous physical properties are assessed with property-specific evaluators rather than a single generic score. This separation also allows multimodal models to be used primarily for high-level perception and localization, leaving physical scoring to explicit acoustic measurements whenever possible.

The following sections instantiate these principles into a structured benchmark construction and an event-grounded evaluation protocol.

## 3 PhyAudioBench: Benchmark Construction

Building on the formulation in Sec. 2, we instantiate **PhyAudioBench**, a hierarchical benchmark for evaluating physical consistency in text-to-audio generation. PhyAudioBench is designed to answer progressively stronger questions about a generated acoustic scene: whether the intended sounds are generated, whether they are organized correctly over time, and whether their acoustic behavior is

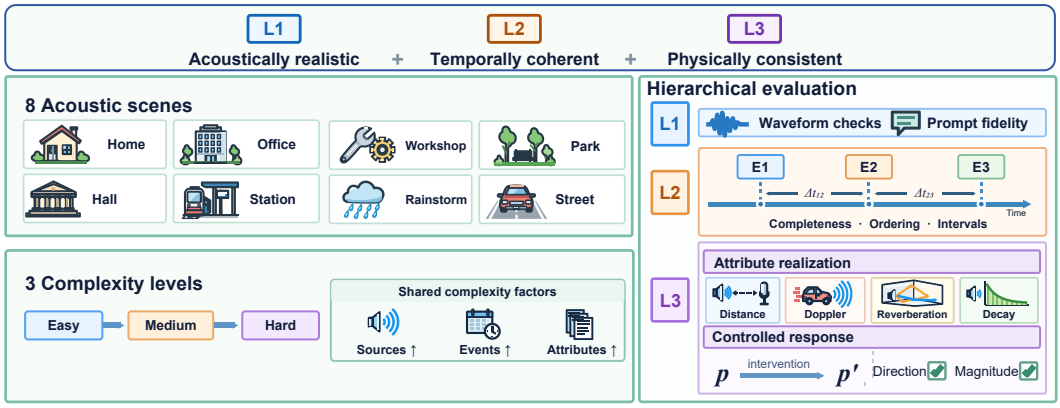



Figure 2: Overview of **PhyAudioBench**. The benchmark organizes evaluation into three hierarchical levels, from acoustic and prompt fidelity to temporal structure and physical consistency. Structured acoustic specifications associate physical properties with individual events, while minimally modified prompt pairs enable controlled evaluation of physical response.

consistent with the described physical conditions. To support such evaluation, each test instance is associated with a structured specification that explicitly represents acoustic events, temporal relations, physical properties, and controlled interventions.

### 3.1 Hierarchical Evaluation

As illustrated in Fig. 2, PhyAudioBench organizes evaluation into three progressively stronger levels. **Level 1: Acoustic and Prompt Fidelity** evaluates whether the generated waveform is acoustically valid and contains the sound content requested by the prompt. **Level 2: Temporal Structure** evaluates whether the required events are present and organized according to the specified temporal relations. **Level 3: Physical Consistency** further evaluates whether the localized events exhibit acoustic behavior compatible with their physical conditions.

The hierarchy reflects the dependency discussed in Sec. 2, rather than merely grouping heterogeneous metrics. A physical property cannot be meaningfully evaluated if its target event is absent, and event-specific physical measurement requires that the corresponding event first be temporally grounded. The three levels therefore distinguish failures in basic generation, temporal composition, and physical behavior instead of collapsing them into a single notion of generation quality.

At Level 3, we consider two complementary forms of physical evaluation. *Attribute realization* evaluates the acoustic evidence for a physical property within an individual generation, whereas *controlled response* evaluates whether that evidence changes consistently when the corresponding physical condition is intervened upon. The former determines whether a physical characteristic is expressed in the generated audio; the latter probes whether the generator is actually responsive to the condition that should govern it.

### 3.2 Structured Benchmark Specification

A central requirement of PhyAudioBench is that the factor being evaluated can be identified without unnecessarily entangling unrelated generation requirements. Following the principle of disentangled evaluation in generative benchmarks (Huang et al., 2024), we construct the benchmark synthetically at the *prompt and specification* level. Prompts, event relations, physical properties, targets, and interventions are generated under constrained construction guidelines, while all evaluated audio clips are produced by the audio generation models.

Each instance is composed of semantically identifiable acoustic events. Individual events are kept atomic so that they can be reliably localized and evaluated, while more complex instances combine multiple events or sound sources. When temporal structure is involved, the specification explicitly defines relations among events as an ordered sequence $E_1 \rightarrow E_2 \rightarrow \cdots \rightarrow E_n$. Counts, ordering,

and temporal intervals are stated explicitly whenever applicable, providing observable targets for subsequent evaluation.

PhyAudioBench covers eight representative acoustic scenes: *home*, *office*, *hall*, *station*, *workshop*, *park*, *rainstorm*, and *street*. These scenes span indoor and outdoor environments, different acoustic spaces, human and mechanical activities, and natural sound sources. Each scene is instantiated at three complexity levels (easy, medium, and hard) by progressively increasing the numbers of distinct sound sources, acoustic events, and annotated physical properties. Importantly, individual events remain semantically atomic as complexity increases, so that greater compositional difficulty does not introduce additional coupling within the evaluation target.

Also, we incorporate seven physical properties: *distance*, *Doppler effect*, *reverberation*, *decay*, *resonance*, *echo*, and *occlusion*. Physical properties are attached to individual acoustic events. Each annotated phenomenon is therefore represented by a `property_id` and a `target_id`, where the former identifies the property to be evaluated and the latter identifies the event to which it applies.

### 3.3 Controlled Physical Interventions

PhyAudioBench constructs controlled prompt pairs $(p, p')$ to operationalize the response-based evaluation introduced in Sec. 2.1. The original prompt $p$ defines a reference acoustic scene, while $p'$ changes one targeted physical condition and preserves the sound identity, event structure, temporal relations, and remaining semantic content as much as possible. This *minimal-intervention* principle is intended to reduce unrelated differences between the two conditioning inputs, making the resulting acoustic response more directly attributable to the physical factor being tested.

Each perturbed instance contains a `pair_id` linking it to its reference prompt and an `intervention_id` specifying the physical intervention. The corresponding intervention specification defines how the relevant physical attributes are expected to respond. An attribute may be expected to increase or decrease under the intervention, while other annotated attributes may remain invariant.

The resulting benchmark contains **240 structured prompts** across **8 acoustic scenes** and **3 complexity levels**, covering **7 physical properties** and **12 controlled intervention types**. Because scene content, event structure, physical annotations, and interventions are represented explicitly, the same specification can be connected to different downstream measurements without embedding metric-specific logic into each prompt. Sec. 4 describes how PhyAudioBench grounds these specifications to generated audio and computes the three levels of evaluation.

## 4 Evaluation Protocol

Given a generated waveform and its structured specification, PhyAudioBench follows a simple principle: *use multimodal models to observe, and explicit evaluators to measure*. A multimodal audio model extracts observations such as event identity and temporal location, as it requires high-level understanding, while signal quality and physical properties are assessed using deterministic parsers or dedicated property-specific evaluators.

### 4.1 Shared Perception and Event Grounding

PhyAudioBench performs a structured observation for each generated sample. For a waveform $x$, the observer produces

$$\mathcal{O}(x) = \{\mathbf{m}, i, c, u, p, \mathbf{e}\}, \tag{3}$$

where $\mathbf{m}$ records the detection states of the required main sounds; $i$, $c$, $u$, and $p$ describe irrelevant-sound, count, continuity, and periodicity violations when applicable; and $\mathbf{e}$ contains event-level detections and temporal locations. For event $j$, we represent

$$\mathbf{e}_j = \left(y_j, \left\{[s_{j1}, e_{j1}], \ldots, [s_{jK_j}, e_{jK_j}]\right\}\right), \tag{4}$$

where $y_j$ denotes whether the event is detected, uncertain, or absent, and the interval set contains its estimated occurrences. This representation naturally supports repeated events through multiple disjoint temporal windows.

Table 1: Overview of the metrics in PhyAudioBench.

| Level | Metrics |
|---|---|
| **L1** | **M101** Duration; **M102** Sample rate; **M103** Background noise; **M104** Distortion; **M105** Signal continuity; **M106** Main-sound presence; **M107** Irrelevant-sound suppression; **M108** Count correctness; **M109** Continuity compliance; **M110** Periodicity compliance |
| **L2** | **M201** Event completeness; **M202** Event ordering; **M203** Interval reasonableness |
| **L3** | **M301** Attribute realization; **M302** Intervention response |

Importantly, $\mathcal{O}(x)$ is a shared intermediate representation. Its semantic fields support prompt-fidelity evaluation at Level 1, while $\mathbf{e}$ supports event completeness and temporal reasoning at Level 2. For Level 3, the same event locations are combined with the structured annotation $(\texttt{property_id}, \texttt{target_id})$ introduced in Sec. 3.2. The `target_id` identifies the relevant event region and the `property_id` selects the corresponding physical evaluator:

$$(\texttt{property_id}, \texttt{target_id}) \rightarrow \hat{T}_{\text{target}} \rightarrow f_a \rightarrow u_a, \tag{5}$$

where $\hat{T}_{\text{target}}$ denotes the localized acoustic region, $f_a$ is the evaluator for physical property $a$, and $u_a$ is its raw measurement. Thus, multimodal perception determines *where* to measure, rather than directly deciding whether the resulting audio is physically correct. We use **Qwen3.5-omni-plus** as the shared multimodal observer in all experiments. The complete observer instruction is provided in Appendix B.3.

### 4.2 Hierarchical Metrics

PhyAudioBench instantiates the three-level hierarchy from Sec. 3.1 with 15 metrics. All metric scores are normalized to $[0, 1]$, with larger values indicating better compliance. Metrics that are not applicable to a prompt are marked as N/A and excluded from the corresponding macro-average.

**Level 1: Acoustic and Prompt Fidelity.** M101–M105 operate directly on the generated waveform and assess basic generation and signal quality. M106–M110 instead use the shared observation $\mathcal{O}(x)$ to determine whether the requested sounds are present and whether explicit requirements on irrelevant sounds, counts, continuity, or periodicity are satisfied. Conditional metrics such as count or periodicity are activated only when the corresponding requirement is explicitly specified by the prompt.

**Level 2: Temporal Structure.** Level 2 reuses the event observations $\mathbf{e}$. Given the event relations specified by the benchmark, M201 evaluates whether all required events are present, M202 whether they occur in the correct order, and M203 whether explicitly specified temporal intervals are reasonable.

**Level 3: Physical Consistency.** Level 3 contains the two complementary forms of physical evaluation introduced in Sec. 3.1. Both use the same property-specific evaluator $f_a$, but differ in how its output is interpreted.

For each physical attribute, we calibrate the evaluator using negative and positive reference responses, $\mu_a^-$ and $\mu_a^+$. M301 linearly normalizes the absolute measurement between these references, while M302 scores the intervention-induced difference after aligning it with the expected direction and normalizing its magnitude by the corresponding reference gap. Full definitions are given in Appendix C.3.

### 4.3 Physics-Aware Evaluators

The Level 3 formulation requires measurements that are specific to the physical phenomenon under evaluation. We therefore implement seven signal-based evaluators

$$\begin{aligned} &f_a(x) = s_a(x) \in [0, 1], \\ &\text{for } a \in \{\text{distance}, \text{Doppler}, \text{reverberation}, \text{decay}, \text{resonance}, \text{echo}, \text{occlusion}\}. \end{aligned} \tag{6}$$

Each evaluator is designed around interpretable acoustic evidence. The details are provided in D.

**Decay and reverberation.** Decay measures post-excitation energy reduction using decay rate, dynamic range, and fit quality. Reverberation additionally captures late energy and reverberation time, with an activity check for a clear excitation–tail structure.

**Distance and occlusion.** Distance combines overall level, high-frequency attenuation, and late-energy behavior. Occlusion uses similar cues but emphasizes preferential high-frequency suppression to reduce ambiguity from source amplitude.

**Doppler, echo, and resonance.** Doppler tracks time-varying spectral-centroid trajectories. Echo measures delayed repetition through onset-strength autocorrelation. Resonance captures free ringing and sustained spectral modes through prominence, concentration, stability, and ringing behavior.

These evaluators are intentionally modular. The benchmark formulation depends only on the availability of a property measurement $f_a$; a stronger evaluator for the same physical property can therefore replace the current implementation without changing event routing or the controlled-intervention formulation.

### 4.4 Evaluator Validation

Because Level 3 conclusions depend on the physical measurements above, we validate whether each evaluator responds to its intended acoustic property. For properties admitting a meaningful positive–negative comparison, positive examples are constructed to explicitly exhibit the target effect, while negative examples are drawn from a shared collection of acoustically diverse clips in which that effect is not explicitly requested. The negative collection is not assumed to be physically free of the property.

We report the difference between positive and negative mean scores,

$$\Delta\mu_a = \mu_a^+ - \mu_a^-, \tag{7}$$

together with ROC–AUC. As shown in Table 2, all five evaluators assessed under this protocol produce higher average scores for their corresponding positive examples. Decay provides the clearest separation, with $\Delta\mu = 0.458$ and AUC $0.916$. Occlusion, resonance, and reverberation yield AUC values between $0.664$ and $0.681$, showing more moderate but consistent separation.

Table 2: Physics-aware evaluator validation. Echo is excluded due to insufficiently diverse positive examples.

| Evaluator | $N^+$ | $N^-$ | $\mu^+$ | $\mu^-$ | $\Delta\mu$ | AUC |
|---|---|---|---|---|---|---|
| Decay | 36 | 170 | 0.666 | 0.208 | 0.458 | 0.916 |
| Doppler | 40 | 170 | 0.634 | 0.330 | 0.304 | 0.842 |
| Occlusion | 32 | 170 | 0.712 | 0.569 | 0.143 | 0.664 |
| Resonance | 32 | 170 | 0.578 | 0.427 | 0.151 | 0.681 |
| Reverberation | 26 | 170 | 0.707 | 0.581 | 0.126 | 0.668 |

Distance requires a different validation protocol because every audible event necessarily occurs at some distance. For echo, the evaluator is defined analytically because the available samples are too homogeneous to support reliable binary discrimination. Detailed evaluator definitions, validation-set construction, and the complete analyses of distance and echo are provided in Appendix E.

## 5 Benchmarking Current Audio Generators

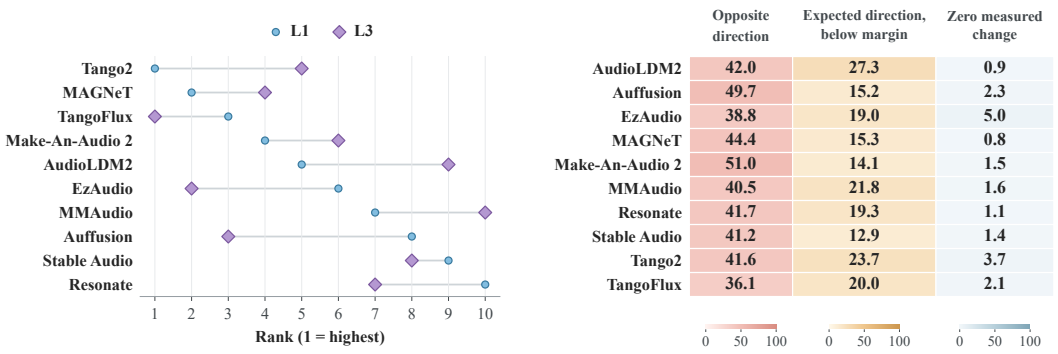
We use PhyAudioBench to evaluate ten representative audio generation models spanning diffusion, flow-matching, diffusion-Transformer, and non-autoregressive architectures. Rather than reducing evaluation to an aggregate model ranking, we examine whether strong performance at lower evaluation levels carries over from acoustic fidelity to temporal structure and ultimately to physical consistency.

### 5.1 Experimental Setup

We evaluate **TangoFlux** (Hung et al., 2024), **Stable Audio Open 1.0** (Evans et al., 2024), **Auffusion** (Xue et al., 2024), **MAGNeT** (Ziv et al., 2024), **EzAudio** (Hai et al., 2024), **Resonate** (Li et al., 2026), **MMAudio** (Cheng et al., 2025), **Tango 2** (Majumder et al., 2024), **AudioLDM 2** (Liu et al., 2024), and **Make-An-Audio 2** (Huang et al., 2023b). Together, these models cover a broad range of architectures, parameter scales, training objectives, and generation focuses.

Table 3: Performance of current audio generators on PhyAudioBench.

| Model | L1 | L2 | L3 | Overall |
|---|---|---|---|---|
| AudioLDM 2 (Liu et al., 2024) | 0.8481 | 0.4044 | 0.1823 | 0.4783 |
| Make-An-Audio 2 (Huang et al., 2023b) | 0.8836 | 0.5930 | 0.2636 | 0.5801 |
| Auffusion (Xue et al., 2024) | 0.8346 | 0.6474 | 0.2740 | 0.5853 |
| EzAudio (Hai et al., 2024) | 0.8474 | 0.5672 | 0.2789 | 0.5645 |
| MAGNeT (Ziv et al., 2024) | 0.8916 | 0.5575 | 0.2728 | 0.5740 |
| MMAudio (Cheng et al., 2025) | 0.8400 | 0.3282 | 0.1583 | 0.4422 |
| Stable Audio Open 1.0 (Evans et al., 2024) | 0.8006 | 0.4479 | 0.2366 | 0.4950 |
| Tango2 (Majumder et al., 2024) | **0.9028** | 0.6178 | 0.2694 | 0.5966 |
| TangoFlux (Hung et al., 2024) | 0.8878 | **0.7183** | **0.3361** | **0.6474** |
| Resonate (Li et al., 2026) | 0.7979 | 0.5103 | 0.2431 | 0.5171 |




(a) L1 vs. L3 rankings.

(b) M302 failure modes.

Figure 3: Diagnosing current audio generation models with PhyAudioBench. (a) Model rankings change substantially between L1 and L3. (b) Incorrect response direction is a major source of controlled-intervention failure.

Each model is evaluated using a fixed model-specific inference configuration that is held constant across all benchmark instances and random seeds. We use a base seed of `20260915` with seed IDs $\{0, 1, 2\}$. The model configurations used in our evaluation, including model size, architecture, sampling rate, and generation steps, are reported in Appendix G. Our results therefore characterize model behavior under the reported configurations rather than the maximally tuned performance of each individual model. Complete per-metric results are provided in Appendix F.

### 5.2 Physical Consistency Scores Are Consistently Lowest

A consistent pattern emerges across all evaluated models: performance decreases as the benchmark moves from acoustic and prompt fidelity to temporal structure and physical consistency. Every model satisfies

$$\text{L1} > \text{L2} > \text{L3}.$$

As shown in Table 3, all ten models exhibit the same ordering, with L1 scores exceeding L2 scores and L2 scores exceeding L3 scores.

This systematic separation exposes a capability gap that is largely invisible under conventional acoustic evaluation. Current generators are comparatively successful at producing plausible waveforms and the requested sound content, but their performance deteriorates once evaluation requires correct temporal composition and, especially, physically consistent acoustic behavior. Models with similar L1 scores can exhibit substantially different L3 performance, indicating that acoustic plausibility alone does not characterize whether a model respects the physical conditions described by the prompt.

### 5.3 Model Rankings Shift from Acoustic to Physical Evaluation

The hierarchy reveals a substantial change in the relative behavior of different models. The three highest-ranked models at L1 are Tango2, MAGNeT, and TangoFlux, whereas the three highest-ranked models at L3 are TangoFlux, EzAudio, and Auffusion. Only TangoFlux appears in both groups. In fact, all ten evaluated models change rank between L1 and L3, with a mean absolute rank shift of $3.0$ positions. The Spearman rank correlation between L1 and L3 is only $0.3697$, compared with $0.4182$ between L1 and L2 and $0.8667$ between L2 and L3 (Fig. 3a).

These ranking changes provide information beyond the raw level-score gap alone. The modest L1–L3 rank correlation indicates that physical evaluation captures aspects of model behavior that are not reflected by L1 ranking alone. Consequently, selecting models solely according to acoustic and prompt fidelity can lead to a substantially different ordering from evaluating their behavior under physical constraints.

Physical-consistency performance also varies substantially across acoustic scenes. Although TangoFlux leads five of the eight scenes, three other models rank first in at least one scene, and individual models can exhibit substantial cross-scene rank shifts. This variation suggests that physical capability is not uniformly expressed across acoustic environments and event compositions. Complete scene-wise rankings are provided in Appendix F.2.

### 5.4 Physical Responses Often Change in the Wrong Direction

The controlled interventions provide a more diagnostic view of why physical consistency remains difficult. A low M302 score can arise because a measurable response changes below margin, does not measurably change, or changes in the *opposite* direction from the expected physical response. property measurement cannot be obtained for one side of the comparison. Figure 3b decomposes these failure modes across models.

Among the measurable response categories, incorrect direction is the most prominent failure mode. Across all ten models, opposite-direction changes account for $36.1\%$–$51.0\%$ of M302 failure cases. By comparison, changes in the expected direction but below the scoring margin account for $12.9\%$–$27.3\%$, while zero measured change accounts for only $0.8\%$–$5.0\%$. Thus, current models do not merely produce physical responses that are too weak; when a measurable response is produced, it frequently changes the relevant acoustic evidence in a direction inconsistent with the controlled intervention.

This observation sharpens the distinction between *acoustic plausibility* and *conditional physical behavior*. A model may have learned plausible local acoustic patterns like attenuation, reverberation, or spectral variation without reliably associating those patterns with the physical conditions that should control them. PhyAudioBench makes this failure visible by evaluating how acoustic evidence changes under minimally modified physical conditions.

## 6 Conclusion

We introduced **PhyAudioBench**, a benchmark for evaluating physical consistency in audio generation beyond acoustic plausibility and semantic correctness. PhyAudioBench combines hierarchical evaluation with controlled physical interventions, event-level grounding, and property-specific acoustic measurements to assess not only whether the intended sounds are generated, but also whether they respond consistently to changes in the described physical conditions. Evaluating ten representative audio generation models reveals a substantial gap between acoustic fidelity and physical consistency: model rankings change markedly across evaluation levels, and controlled interventions frequently induce acoustic changes in the wrong physical direction. These results highlight physical consistency as a distinct and still underdeveloped capability of current models, and establish PhyAudioBench as a systematic testbed for measuring this gap.

### AI USE STATEMENT

Generative AI tools were used in this work to assist with research ideation and refinement, to draft and polish portions of the manuscript, and to support the construction of synthetic benchmark data. In particular, AI assistance was used to help formulate and refine research ideas, improve the clarity and presentation of the paper, and generate or refine structured prompts and specifications under author-defined constraints for synthetic benchmark construction. All AI-assisted research decisions, generated benchmark content, and manuscript text were reviewed and verified by the authors. The authors take full responsibility for the methodology, experiments, results, claims, and final content of the paper.

### ETHICS STATEMENT

This work focuses on evaluating audio generation models and does not involve human subjects, private data, or personally identifiable information. Benchmark prompts and specifications are synthetically constructed. While improved evaluation may indirectly support more capable audio generation systems, our work does not introduce new generation capabilities. We follow the applicable licenses and usage terms of all models and resources used in the study.

### REPRODUCIBILITY STATEMENT

The appendix provides the complete inference configurations for all evaluated audio generation models, including model variants, sampling rates, decoding or diffusion steps, and seed settings. We also report the multimodal audio model used as the shared observer, together with its prompting and structured-output protocol. Full definitions of the downstream evaluators, physical-property measurements, normalization procedures, intervention-response scoring rules, and attribute-specific reference calibration are provided in the appendix. In addition, we document the benchmark construction procedure, including scene categories, complexity levels, event composition rules, physical-property annotations, intervention specifications, and prompt-generation constraints. These details are intended to enable reproduction of both the generated evaluation set and the complete scoring pipeline.

## REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28901–28911, 2025.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yusheng Dai, Zehua Chen, Yuxuan Jiang, Baolong Gao, Qiuhong Ke, Jianfei Cai, and Jun Zhu. Omni2sound: Towards unified video-text-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740, 2020. doi: 10.1109/ICASSP40776.2020.9052990.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. Ezaudio: Enhancing text-to-audio generation with efficient diffusion transformer. *arXiv preprint arXiv:2409.10819*, 2024.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. Vabench: A comprehensive benchmark for audio-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-yi Lee. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. *arXiv preprint arXiv:2309.09510*, 2023a.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of NAACL-HLT*, pp. 119–132, 2019. doi: 10.18653/v1/N19-1011.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *International Conference on Learning Representations*, 2023.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xiquan Li, Junxi Liu, Wenxi Chen, Haina Zhu, Ziyang Ma, and Xie Chen. Resonate: Reinforcing text-to-audio generation via online feedback from large audio language models. *arXiv preprint arXiv:2603.11661*, 2026.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. doi: 10.1109/TASLP.2024.3399607.

Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, Leilei Sun, Dahua Lin, and Jiaqi Wang. STAR-Bench: Probing Deep Spatio-Temporal Reasoning as Audio 4D Intelligence. In *International Conference on Learning Representations*, 2026.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024. doi: 10.1145/3664647.3681688.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8293–8316, 2024. doi: 10.18653/v1/2024.naacl-long.459.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

Yuchen Sun, Qian Yang, Jun Wang, Detai Xin, Guoqiao Yu, Guanglu Wan, and Qi Jia. Prism-bench: An audio-centric diagnostic benchmark for text-to-audio-video generation, 2026.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: A unified framework for anything-to-audio generation. In *International Conference on Learning Representations*, 2026.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. Hear: Holistic evaluation of audio representations. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176, pp. 125–145. PMLR, 2022.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*, 2024.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Hui Wang, Cheng Liu, Junyang Chen, Haoze Liu, Yuhang Jia, Shiwan Zhao, Jiaming Zhou, Haoqin Sun, Hui Bu, and Yong Qin. Tta-bench: A comprehensive benchmark for evaluating text-to-audio models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33512–33520, 2026. doi: 10.1609/aaai.v40i39.40639.

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *arXiv preprint arXiv:2401.01044*, 2024.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. Superb: Speech processing universal performance benchmark. In *Interspeech*, 2021.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, et al. Marble: Music audio representation benchmark for universal evaluation. *arXiv preprint arXiv:2306.10548*, 2023.

Qian Zhang, Yuqin Cao, Yixuan Gao, and Xiongkuo Min. Vidaudio-bench: Benchmarking v2a and vt2a generation across four audio categories. *arXiv preprint arXiv:2604.10542*, 2026.

Ziwei Zhou, Zeyuan Lai, Rui Wang, Yifan Yang, Zhen Xing, Yuqing Yang, Qi Dai, Lili Qiu, and Chong Luo. Avgen-bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation. *arXiv preprint arXiv:2604.08540*, 2026.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer. In *International Conference on Learning Representations*, 2024.

# A Related Work

## A.1 Audio Generation Models

Audio generation has progressed rapidly from domain-specific synthesis toward increasingly general and controllable generation. Recent text-to-audio (T2A) models can synthesize diverse environmental sounds, human activities, and everyday acoustic events directly from natural-language descriptions. Speech generation models include VALL-E (Wang et al., 2023) and Voicebox (Le et al., 2023), while music generation models include MusicLM (Agostinelli et al., 2023) and MusicGen (Copet et al., 2023). Early models like AudioGen (Kreuk et al., 2023) and AudioLM (Borsos et al., 2022) both operate on discrete audio tokens. Representative systems such as AudioLDM 2 (Liu et al., 2024) formulate general audio generation in a latent diffusion framework, while Stable Audio Open (Evans et al., 2024) adopts a diffusion-transformer architecture for open-weight audio generation, and TangoFlux (Hung et al., 2024) explores flow matching for efficient and high-quality text-conditioned synthesis. Beyond improving perceptual quality, recent models increasingly emphasize semantic alignment, temporal organization, and compositional generation, enabling more complex prompts containing multiple sound events and temporal relations.

Meanwhile, audio generation is becoming increasingly multimodal. Video-to-audio (V2A) models generate sounds aligned with visible events, while video-text-to-audio (VT2A) systems further incorporate textual instructions to complement visual conditioning. Unified models such as Omni2Sound (Dai et al., 2026) support T2A, V2A, and VT2A within a single framework. More general anything-to-audio systems such as AudioX (Tian et al., 2026) extend conditioning to text, video, image, and audio, supporting both generation and editing under flexible multimodal inputs. These advances substantially broaden the range and complexity of acoustic scenes that modern generators can model, creating a corresponding need for evaluation protocols that go beyond perceptual quality and coarse semantic alignment.

## A.2 Audio Evaluation

Audio evaluation has developed along different axes across speech, music, and general audio. AudioSet (Gemmeke et al., 2017) provides a hierarchical ontology and a large human-labeled corpus of audio events for audio-event recognition and detection. AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) provide audio-caption pairs for audio captioning. For speech, benchmarks such as SUPERB (Yang et al., 2021) and Dynamic-SUPERB (Huang et al., 2023a) organize evaluation around well-defined linguistic and paralinguistic capabilities, including content recognition, speaker characteristics, and emotion-related information. AudioBench (Wang et al., 2024) evaluates audio-language models across speech understanding, audio scene understanding, and paralinguistic understanding tasks. Music benchmarks such as MARBLE (Yuan et al., 2023) and MusicBench (Melechovsky et al., 2024) instead focus on music-specific representations and perceptual properties, including aspects related to rhythm, instrumentation, and musical structure. For general audio, evaluation has gradually moved from coarse acoustic perception toward more fine-grained understanding and generation assessment. HEAR (Turian et al., 2022) evaluates audio representations across speech, environmental sound, and music, while AIR-Bench (Yang et al., 2024) and MMAU (Sakshi et al., 2024) target broader audio understanding and reasoning. STAR-Bench (Liu et al., 2026), for example, probes fine-grained acoustic perception together with temporal and spatial reasoning, while T2A-Bench (Tian et al., 2026) evaluates text-to-audio generation through sound category, event count, temporal ordering, and timestamp consistency. Together, these benchmarks substantially improve our ability to measure whether generated audio contains the intended events and whether those events are organized correctly in time, but physical acoustic behavior remains much less explicitly characterized. For visually conditioned audio generation, VidAudio-Bench (Zhang et al., 2026) evaluates V2A/VT2A systems with task-specific metrics, while VABench (Hua et al., 2026) and AVGen-Bench (Zhou et al., 2026) assess joint audio–video generation.

Recent audio benchmarks also increasingly employ multimodal large language models as flexible evaluators for open-ended acoustic properties. For example, PRISM-Bench (Sun et al., 2026) uses an MLLM-based judge to assess a broad set of fine-grained perceptual criteria, while TTA-Bench (Wang et al., 2026) combines automatic metrics with extensive expert and user annotations across multiple evaluation dimensions. Although these approaches substantially broaden evaluation coverage, they expose two practical limitations. First, directly using an MLLM for final scoring makes the evaluation dependent on the perception, reasoning, and calibration of the selected judge. Second, extending a benchmark to new capabilities often requires additional task-specific prompts, annotations, scoring rules, or human evaluation. Our work instead emphasizes structured observation, explicit property-specific measurement, and reusable evaluation specifications, aiming to reduce both dependence on unrestricted learned judges and the manual cost of extending evaluation to new physical behaviors.

# B Benchmark Specification and Annotation Details

This section provides additional details of the structured benchmark specification and the shared multimodal observation used by PhyAudioBench. The benchmark construction is described in Sec. 3, while the evaluation pipeline is described in Sec. 4.

## B.1 Representative Benchmark Instance

Each benchmark instance associates a natural-language prompt with structured annotations describing its acoustic events, physical properties, and, when applicable, a controlled intervention. A representative intervention pair from the *office* scene is

> **Reference prompt $p$:** "A phone rings once nearby."
> **Intervened prompt $p'$:** "A phone rings once far away."
> **Intervention:** `increase_distance`
> **Target property:** `distance`
> **Target event:** `phone_ringing`

The two prompts preserve the sound identity and event structure while modifying the source–receiver distance. The corresponding structured annotation links the physical property to its acoustic target through `property_id` and `target_id`, and connects the intervened instance to its reference using `pair_id` and `intervention_id`. The intervention specification further records the expected direction of change for the affected physical properties.

## B.2 Structured Multimodal Observation

As described in Sec. 4.1, a single multimodal observation is shared by Level 1 semantic metrics and Level 2 temporal metrics. For a generated waveform $x$, the observation has the form

$$\mathcal{O}(x) = \{\mathbf{m}, i, c, u, p, \mathbf{e}\}\,. \tag{8}$$

The fields have the following semantics: $\mathbf{m}$ records the perceptual presence of the required main sounds; $i$ indicates whether a prominent irrelevant sound is present; $c$ indicates an explicit count mismatch; $u$ indicates a violation of an explicit continuity requirement; $p$ indicates a violation of an explicit periodicity requirement; and $\mathbf{e}$ contains event-level detection states and temporal intervals.

A representative observation is

```
{
  "m": [
    ["basketball impacts", "d"],
    ["referee whistle", "d"]
  ],
  "i": "n",
  "c": "n",
```

```
  "u": "n",
  "p": "n",
  "e": [
    ["d", [[3.3, 4.3]]],
    ["d", [[6.0, 7.0], [8.1, 9.2]]]
  ]
}
```

Here, `d`, `u`, and `x` denote *detected*, *uncertain*, and *not detected*, respectively. The event intervals are approximate temporal regions in seconds and can contain multiple disjoint windows when an event occurs repeatedly. These structured observations are subsequently converted into metric scores through the deterministic rules described in Appendix C.

### B.3 Full Multimodal Observer Prompt

We use the following instruction for the shared multimodal observer.

```
You are a strict audio benchmark annotator. Analyze the supplied audio
and the reference prompt once, then return one JSON object containing
both L1 prompt fidelity observations and L2 event observations.

General rules

1. Judge only from audible acoustic evidence in the supplied audio.
2. Use "u" when evidence is genuinely partial, masked, ambiguous, or too
weak for a reliable yes/no decision.
3. Do not assume that a sound is present merely because it appears in the
text.
4. Do not add explanations, Markdown, code fences, comments, confidence
prose, units outside numeric interval values, or any field not defined
below.
5. The response must be valid JSON and must follow the exact flat schema
shown at the end of this instruction.

L1 field: m -- required main audible sounds

Infer every main audible sound required by the reference prompt. Return
one [sound_name,status] pair for every required main sound.

Status values are restricted to:
- "d": clearly detected and acoustically consistent with the required
sound.
- "u": uncertain; there is partial, masked, ambiguous, or insufficient
evidence.
- "x": not detected; no meaningful evidence of the required sound is
audible.

Use concise audible names, for example "running footsteps" rather than
"running", and "basketball impacts" rather than "basketball". Do not omit
a required main sound. Do not mark a merely similar or semantically
related sound as detected.

L1 field: i -- obvious irrelevant sound

Return exactly one JSON string:
- "n": no clear, prominent and unrelated sound is audible.
- "y:<sound>": a clear and prominent sound not requested or reasonably
implied by the prompt is audible.

Ignore weak background noise, reverberation, low-level recording
artifacts and sounds whose identity is uncertain. If several irrelevant
sounds exist, report the most prominent one.

L1 field: c -- explicit count error
```

```
Evaluate this field only when the prompt explicitly specifies a count,
such as "once", "twice", "three times", "one hit", or "three hits".

Return exactly one JSON string:
- "n": no explicit count exists, the observed count is consistent, or no
clear significant mismatch can be established.
- "y:<expected>:<observed>": the number of clearly distinguishable
occurrences differs from the explicit required count.

Keep the whole value inside one pair of JSON quotation marks. If expected
and observed are equal, return "n", never a y-value.

L1 field: u -- explicit continuity error

Evaluate only when the prompt explicitly requires a sound to be
continuous, continuously present, or sustained.

Return:
- "n": not applicable, reasonably continuous, or no clear violation.
- "y": the required sound clearly stops, becomes intermittent, or is
absent for a significant part of the required span.

Natural loudness fluctuations alone are not a continuity error.

L1 field: p -- explicit periodicity error

Evaluate only when the prompt explicitly requires periodic repetition or
regular intervals.

Return:
- "n": not applicable, reasonably periodic, or no clear violation.
- "y": the requested periodic pattern is clearly missing, highly
irregular, or significantly inconsistent.

Do not require mechanically perfect timing.

L2 field: e -- YAML subject event detection and time localization

The user supplies a list named "YAML subjects in exact required order".
These are controlled observable event labels extracted only from
relations[*].subject.

For every supplied subject, return exactly one [status,intervals] item,
in the same order. Do not add, remove, rename, merge, split, or reorder
subjects.

Event status values are restricted to "d", "u", and "x":
- "d": the specific subject event is clearly detected.
- "u": the specific subject event is uncertain.
- "x": the specific subject event is not detected.

The intervals value is a JSON list of approximate
[start_second,end_second] ranges:
- For "d", normally provide at least one usable interval.
- For "u", provide candidate intervals when reasonably localizable,
otherwise [].
- For "x", always return [].
- For a repeated event, provide one interval for each clearly
distinguishable occurrence.
- For a continuous event, normally provide one broad interval covering
its audible span.
- Keep intervals ordered by start time.
- Keep all endpoints inside [0,audio_duration].
- Use coarse decimal seconds. Do not fabricate millisecond precision.
- Each interval must satisfy start < end.
```

```
Output protocol

Return exactly one JSON object with exactly these six keys:

{"m":[["sound", "d"]], "i":"n", "c":"n", "u":"n", "p":"n", "e":[["d",
[[0.0, 1.0]]]]}

Valid count-mismatch example:

{"m":[["basketball impacts", "d"]], "i":"n", "c":"y:3:2", "u":"n",
"p":"n", "e":[["d", [[1.0, 1.2], [2.0, 2.2]]]]}

Before returning, verify all of the following:
- The JSON is syntactically valid.
- The only top-level keys are m, i, c, u, p and e.
- Every m status is d/u/x.
- Every e status is d/u/x.
- The number of e items exactly equals the number of supplied YAML
subjects.
- No x event contains intervals.
- c is one complete JSON string: n or y:<expected>:<observed>.
```

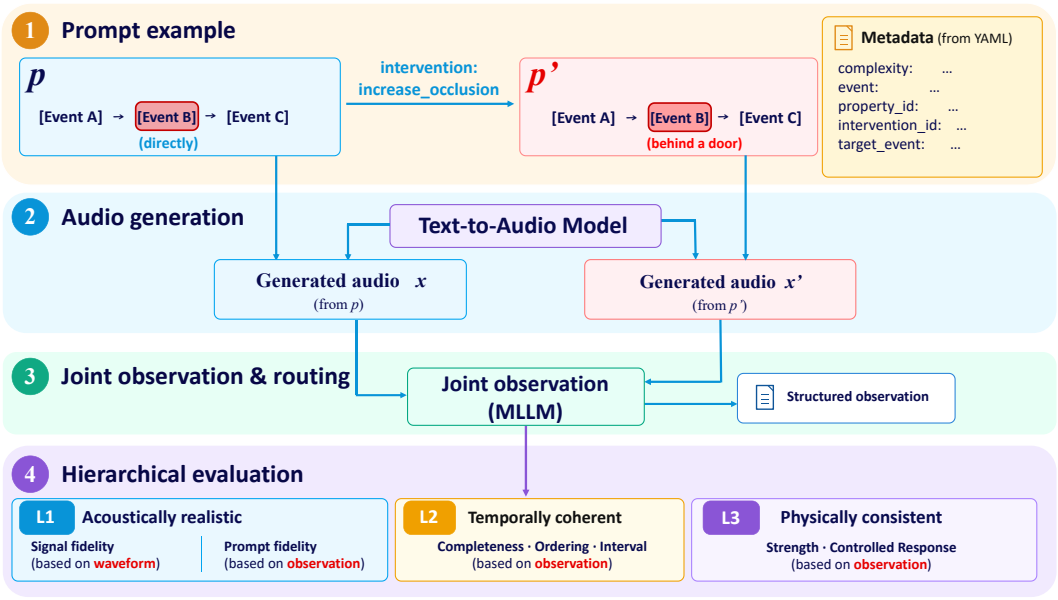



Figure 4: Evaluation pipeline of **PhyAudioBench**. A shared multimodal observer identifies audible events and their temporal locations. The resulting event grounding is reused by prompt- and temporal-consistency metrics, and combined with structured `property_id`–`target_id` annotations to route localized event regions to physics-aware evaluators. Controlled intervention metadata further connects paired generations for response-based evaluation.

## C Detailed Metric Definitions

We provide the complete definitions of the metrics summarized in Sec. 4.2. Let a generated waveform be $x = \{x_t\}_{t=1}^{T}$, let $f_s$ denote its sampling rate, and let $p$ denote its structured prompt specification. All metric scores are normalized to $[0, 1]$, where larger values indicate better compliance. Metrics that are not applicable to a benchmark instance are marked as N/A and excluded from the corresponding aggregation.

### C.1 Level 1: Acoustic and Prompt Fidelity

Level 1 consists of ten metrics. M101–M105 operate directly on the waveform, whereas M106–M110 are derived from the shared multimodal observation described in Appendix B.2.

**M101: Duration Correctness.** Let $D_p$ denote the expected duration of the benchmark instance and

$$D_x = \frac{N_{\text{frame}}}{f_s} \tag{9}$$

the measured duration of the generated waveform. With an absolute tolerance $\tau_D$, we define

$$S_{\text{M101}} = \mathbb{I}\left[|D_x - D_p| \le \tau_D\right]. \tag{10}$$

The current implementation uses $\tau_D = 0.1\,\text{s}$.

**M102: Sample-Rate Correctness.** Let $f_s^*$ denote the target sampling rate required by the evaluation setting and $f_s$ the sampling rate stored in the generated waveform. The metric is

$$S_{\text{M102}} = \mathbb{I}\left[f_s = f_s^*\right]. \tag{11}$$

This metric is intentionally binary and measures generation-format compliance.

**M103: Background Noise.** M103 measures broadband residual noise in low-energy regions using quantities defined relative to the current waveform. The waveform is divided into short overlapping frames, and $r_i$ denotes the RMS energy of frame $i$. We select the lowest-energy fraction $\mathcal{L}$, corresponding to the bottom $q = 0.2$ quantile, and let $\mathcal{A}$ denote the remaining acoustically active frames. The corresponding RMS levels are

$$R_{\text{low}} = \text{mean}_{i\in\mathcal{L}}\, r_i, \qquad R_{\text{act}} = \text{mean}_{i\in\mathcal{A}}\, r_i, \tag{12}$$

with relative level

$$\rho = \frac{R_{\text{low}}}{R_{\text{act}}}. \tag{13}$$

For a low-energy frame with power spectrum $P_i(k)$ over the analysis band, the spectral flatness is

$$F_i = \frac{\exp\left(\frac{1}{K}\sum_{k=1}^{K}\log(P_i(k)+\epsilon)\right)}{\frac{1}{K}\sum_{k=1}^{K}P_i(k)+\epsilon}. \tag{14}$$

The low-energy spectral flatness is

$$F_{\text{low}} = \text{mean}_{i\in\mathcal{L}}\, F_i. \tag{15}$$

Digitally silent frames are excluded from this estimate. The final score is

$$S_{\text{M103}} = 1 - 500\cdot\rho F_{\text{low}}. \tag{16}$$

A substantial penalty therefore requires the low-energy regions to be both relatively energetic and spectrally flat, which is characteristic of broadband residual noise. The current implementation uses $50\,\text{ms}$ frames, a $25\,\text{ms}$ hop, and evaluates spectral flatness over approximately 50–16 000 Hz.

**M104: Distortion.** M104 targets short-duration waveform abnormalities while avoiding penalization of legitimate acoustic transients. Let

$$C = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}\left[|x_t| \ge \tau_{\text{clip}}\right] \tag{17}$$

denote the clipping ratio. We additionally compute adjacent-sample differences

$$d_t = |x_t - x_{t-1}| \tag{18}$$

and detect sufficiently short, isolated excursions above an adaptive local threshold. Let $A$ denote the normalized rate of these abnormal excursions. The distortion penalty is

$$P_{\text{dist}} = \text{clip}_{[0,1]}\left(w_c C + w_a A\right), \tag{19}$$

and the metric score is

$$S_{\text{M104}} = \exp(-P_{\text{dist}}). \tag{20}$$

The short-duration constraint prevents normal high-energy events, such as impacts or thunder strikes, from being penalized solely because they contain large transients.

**M105: Signal Continuity.** M105 evaluates two technical discontinuities: prolonged near-digital silence and abrupt waveform breaks. The waveform is divided into non-overlapping short-time frames, and a frame is marked as near-silent when its RMS level falls below $\tau_{\text{sil}}$. Only contiguous silent runs exceeding a minimum duration $D_{\min}$ are retained. Let $\mathcal{R}_{\text{sil}}$ denote these accepted runs. The silence ratio is

$$R_{\text{sil}} = \frac{\sum_{r \in \mathcal{R}_{\text{sil}}} D(r)}{D_{\text{audio}}}, \tag{21}$$

where $D(r)$ denotes the duration of silent run $r$ and $D_{\text{audio}}$ is the total waveform duration.

Abrupt discontinuities are detected from

$$d_t = |x_t - x_{t-1}|. \tag{22}$$

A candidate must be both a robust global outlier and substantially larger than the local derivative scale on its two sides. Nearby candidates are merged into a single discontinuity event. Let $N_{\text{break}}$ denote the resulting number of waveform breaks. We define

$$R_{\text{break}} = \frac{N_{\text{break}}}{D_{\text{audio}}}. \tag{23}$$

Short isolated pop-like excursions are handled by M104 rather than M105 to avoid double penalization. The continuity score is

$$S_{\text{M105}} = \exp\left(-w_s R_{\text{sil}} - w_b R_{\text{break}}\right), \tag{24}$$

where $w_s$ and $w_b$ control the contributions of prolonged silence and waveform discontinuities, respectively. The implementation uses $20\,\text{ms}$ non-overlapping frames, a silence threshold of $-80\,\text{dBFS}$, and a minimum silent-run duration of $500\,\text{ms}$.

**Shared semantic observation for M106–M110.** For each expected main sound $j$, the observer returns

$$y_j \in \{\text{detected}, \text{uncertain}, \text{not detected}\}.$$

We map these states to

$$c(y_j) = \begin{cases} 1, & y_j = \text{detected}, \\ 0.5, & y_j = \text{uncertain}, \\ 0, & y_j = \text{not detected}. \end{cases} \tag{25}$$

For M107–M110, the observer instead returns a violation judgment $y_j \in \{\text{“y”}, \text{“n”}\}$. We define

$$c_{\text{err}}(y_j) = \begin{cases} 0, & y_j = \text{“y”}, \\ 1, & y_j = \text{“n”}. \end{cases} \tag{26}$$

**M106: Main-Sound Presence.** For $N_m$ required main sounds,

$$S_{\text{M106}} = \frac{1}{N_m} \sum_{j=1}^{N_m} c(y_j). \tag{27}$$

Uncertain detections therefore receive partial credit.

**M107: Irrelevant-Sound Suppression.** Let $y_{\text{irr}}$ denote whether a prominent sound unrelated to the requested events is detected. The score is

$$S_{\text{M107}} = c_{\text{err}}\left(y_{\text{irr}}\right). \tag{28}$$

**M108: Count Correctness.** For prompts containing an explicit count requirement, let $y_{\text{cnt}}$ denote whether the observed number of occurrences differs from the requested count. We define

$$S_{\text{M108}} = c_{\text{err}}\left(y_{\text{cnt}}\right). \tag{29}$$

If no explicit count is specified, M108 is marked N/A.

**M109: Continuity Compliance.** For prompts explicitly requiring a continuous or sustained sound, $y_{\text{cts}}$ denotes whether an unintended interruption is detected. The score is

$$S_{\text{M109}} = c_{\text{err}}\left(y_{\text{cts}}\right). \tag{30}$$

If no explicit continuity requirement is specified, M109 is marked N/A.

**M110: Periodicity Compliance.** For prompts explicitly requiring periodic repetition or regular intervals, $y_{\text{per}}$ denotes whether the required temporal regularity is violated. We define

$$S_{\text{M110}} = c_{\text{err}}\left(y_{\text{per}}\right). \tag{31}$$

If no explicit periodicity or regular-interval requirement is specified, M110 is marked N/A.

### C.2 Level 2: Temporal Structure

Level 2 evaluates whether the required events are present and temporally organized according to the benchmark specification. For event $E_i$, the multimodal observer produces a detection confidence $c_i$ and one or more temporal intervals

$$\hat{\mathcal{T}}_i = \{[\hat{s}_{i1}, \hat{e}_{i1}], \ldots, [\hat{s}_{iK_i}, \hat{e}_{iK_i}]\}. \tag{32}$$

**M201: Event Completeness.** Each expected event uses the same confidence mapping as M106:

$$c_i = \begin{cases} 1, & \text{detected}, \\ 0.5, & \text{uncertain}, \\ 0, & \text{not detected}. \end{cases} \tag{33}$$

For an event chain containing $N_e$ expected events,

$$S_{\text{M201}} = \frac{1}{N_e} \sum_{i=1}^{N_e} c_i. \tag{34}$$

The estimated event intervals are reused by M202, M203, and Level 3 target localization.

**M202: Event Ordering.** For two consecutive expected events $E_i \to E_{i+1}$, if an event is detected in multiple temporal regions, we use its earliest onset as the representative start time:

$$\hat{s}_i = \min_k \hat{s}_{i,k}. \tag{35}$$

The onset difference is

$$\Delta_i = \hat{s}_{i+1} - \hat{s}_i. \tag{36}$$

We use the soft ordering function

$$g_{\text{ord}}(\Delta_i) = \begin{cases} 0, & \Delta_i \le -1, \\ \dfrac{\Delta_i + 1}{2}, & -1 < \Delta_i < 1, \\ 1, & \Delta_i \ge 1. \end{cases} \tag{37}$$

An event beginning at least $1\,\text{s}$ before its predecessor receives no ordering credit, simultaneous onsets receive $0.5$, and a start at least $1\,\text{s}$ later receives full credit.

The confidence-weighted score of an adjacent pair is

$$S_{\text{ord}}^{(i)} = g_{\text{ord}}(\Delta_i) \min(c_i, c_{i+1}), \tag{38}$$

and M202 averages over all adjacent pairs:

$$S_{\text{M202}} = \frac{1}{N_e - 1} \sum_{i=1}^{N_e - 1} S_{\text{ord}}^{(i)}. \tag{39}$$

**M203: Interval Reasonableness.** M203 is evaluated only for adjacent event relations carrying an explicit interval class. The current benchmark supports `shortly_after` and `later`. For an adjacent pair $E_i \to E_{i+1}$, the inter-event gap is

$$G_i = \hat{s}_{i+1} - \hat{e}_i, \tag{40}$$

where

$$\hat{s}_{i+1} = \min_k \hat{s}_{i+1,k}, \qquad \hat{e}_i = \max_k \hat{e}_{i,k}. \tag{41}$$

For `shortly_after`, we use

$$g_{\text{short}}(G_i) = \begin{cases} 0, & G_i < 0, \\ 1, & 0 \le G_i \le 1, \\ \dfrac{3 - G_i}{2}, & 1 < G_i < 3, \\ 0, & G_i \ge 3. \end{cases} \tag{42}$$

Thus, gaps of at most 1 s receive full credit, after which the score decreases linearly to zero at 3 s. Negative gaps receive zero under this endpoint-to-onset definition.

For `later`, we use

$$g_{\text{later}}(G_i) = \begin{cases} 0, & G_i \le 1, \\ \dfrac{G_i - 1}{2}, & 1 < G_i < 3, \\ 1, & G_i \ge 3. \end{cases} \tag{43}$$

Gaps no greater than 1 s therefore receive no credit, while gaps of at least 3 s receive full credit.

Let $r_i \in \{\texttt{shortly_after}, \texttt{later}\}$ denote the annotated relation. The confidence-weighted pair score is

$$S_{\text{int}}^{(i)} = g_{r_i}(G_i) \min(c_i, c_{i+1}), \tag{44}$$

and M203 averages over the set $\mathcal{R}$ of relations with explicit interval annotations:

$$S_{\text{M203}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} S_{\text{int}}^{(i)}. \tag{45}$$

When no explicit interval relation is present, M203 is marked N/A.

### C.3 Level 3: Physical Consistency

Level 3 uses property-specific evaluators to measure physical acoustic evidence within localized target events. For each physical property

$$a \in \{\text{distance}, \text{doppler}, \text{reverberation}, \text{decay}, \text{resonance}, \text{echo}, \text{occlusion}\},$$

let $f_a$ denote the corresponding evaluator. Given a target interval $W_k$,

$$u_{a,k} = f_a(x[W_k]). \tag{46}$$

If a target occurs in multiple valid intervals, the measurements are combined as

$$u_a = \text{Agg}\left(u_{a,1}, \ldots, u_{a,K}\right), \tag{47}$$

where Agg is the evaluator-specific aggregation rule, typically a mean over valid windows.

**M301: Physical-Attribute Realization.** For each property $a$, let

$$\mu_a^- = \frac{1}{N_a^-} \sum_{j=1}^{N_a^-} f_a(x_j^-), \qquad \mu_a^+ = \frac{1}{N_a^+} \sum_{j=1}^{N_a^+} f_a(x_j^+) \tag{48}$$

denote the mean evaluator responses over the lower- and higher-response calibration reference collections. The raw target measurement is normalized as

$$\tilde{u}_a = \frac{u_a - \mu_a^-}{\mu_a^+ - \mu_a^-}, \tag{49}$$

followed by

$$s_{\mathrm{M301}}^{(a)} = \mathrm{clip}_{[0,1]}\left(\tilde{u}_a\right). \tag{50}$$

If multiple properties are routed to M301,

$$S_{\mathrm{M301}} = \frac{1}{|\mathcal{A}_p|} \sum_{a \in \mathcal{A}_p} s_{\mathrm{M301}}^{(a)}, \tag{51}$$

where $\mathcal{A}_p$ contains all valid attributes assigned to M301.

For instances with an `intervention_id`, $\mathcal{A}_p$ contains only attributes marked as *invariant* by the intervention specification. Attributes expected to change under the intervention are excluded from M301 and evaluated by M302 instead. This avoids evaluating the same annotated property simultaneously through absolute realization and paired response.

**M302: Controlled Intervention Response.** For a controlled pair $(p, p')$ with generated waveforms $(x, x')$, the intervention registry provides an expected direction

$$\delta_a \in \{-1, +1\} \tag{52}$$

for each intervention-affected physical property $a$. Attributes marked as invariant are excluded from M302 and routed to M301.

For each comparable property,

$$u_a = f_a(x), \qquad u'_a = f_a(x'), \tag{53}$$

with raw change

$$\Delta u_a = u'_a - u_a. \tag{54}$$

The change is aligned with the expected direction through

$$\Delta_a^{\mathrm{signed}} = \delta_a \Delta u_a = \delta_a \left(u'_a - u_a\right), \qquad \delta_a \in \{-1, +1\}. \tag{55}$$

Thus, $\Delta_a^{\mathrm{signed}} > 0$ denotes a response in the expected direction, whereas $\Delta_a^{\mathrm{signed}} \leq 0$ denotes either no measurable response or an opposite-direction response.

To normalize the required response magnitude across heterogeneous properties, we use the reference gap

$$G_a = \mu_a^+ - \mu_a^-. \tag{56}$$

The margin for full directional credit is

$$m_a = \rho G_a, \tag{57}$$

with

$$\rho = 0.20. \tag{58}$$

A response equal to $20\%$ of the positive–negative reference gap therefore receives full directional credit.

The per-property M302 score is

$$S_{\mathrm{M302}}^{(a)} = \begin{cases} 0, & \Delta_a^{\mathrm{signed}} \leq 0, \\ \mathrm{clip}_{[0,1]}\left(\dfrac{\Delta_a^{\mathrm{signed}}}{\rho\left(\mu_a^+ - \mu_a^-\right)}\right), & \Delta_a^{\mathrm{signed}} > 0. \end{cases} \tag{59}$$

M302 therefore uses direction as a gate and assigns continuous credit to correctly directed responses according to their magnitude.

If an intervention affects multiple comparable property–target pairs, their scores are macro-averaged:

$$S_{\mathrm{M302}} = \frac{1}{|\mathcal{C}_{\mathcal{I}}|} \sum_{(a,t) \in \mathcal{C}_{\mathcal{I}}} S_{\mathrm{M302}}^{(a,t)}, \tag{60}$$

where $\mathcal{C}_{\mathcal{I}}$ contains changed $(\texttt{property_id}, \texttt{target_id})$ pairs that are present in both $p$ and $p'$ and yield valid raw measurements. The resulting pair score is assigned to the intervened instance $p'$.

### C.4 Score Aggregation

For each metric $m$, let $\mathcal{D}_m$ denote the set of benchmark instances for which that metric is valid and applicable. We first compute the metric-level mean

$$\bar{S}_m = \frac{1}{|\mathcal{D}_m|} \sum_{i \in \mathcal{D}_m} S_m^{(i)}. \tag{61}$$

Entries marked as N/A are therefore excluded from the corresponding metric average rather than treated as failures.

Let

$$\mathcal{M}_1 = \{\mathrm{M101}, \ldots, \mathrm{M110}\}, \qquad \mathcal{M}_2 = \{\mathrm{M201}, \mathrm{M202}, \mathrm{M203}\}.$$

The Level 1 and Level 2 scores are the unweighted macro-averages of their metric-level means:

$$S_{\mathrm{L1}} = \frac{1}{|\mathcal{M}_1|} \sum_{m \in \mathcal{M}_1} \bar{S}_m, \tag{62}$$

and

$$S_{\mathrm{L2}} = \frac{1}{|\mathcal{M}_2|} \sum_{m \in \mathcal{M}_2} \bar{S}_m. \tag{63}$$

Level 3 contains two metrics and is computed as

$$S_{\mathrm{L3}} = \frac{\bar{S}_{\mathrm{M301}} + \bar{S}_{\mathrm{M302}}}{2}. \tag{64}$$

Finally, the overall benchmark score reported in Table 3 is the unweighted mean of the three level scores:

$$S_{\mathrm{Overall}} = \frac{S_{\mathrm{L1}} + S_{\mathrm{L2}} + S_{\mathrm{L3}}}{3}. \tag{65}$$

## D Physics-Aware Evaluator Details

This section provides the complete definitions of the seven physics-aware evaluators summarized in Sec. 4.3. Each evaluator operates on a localized target region and extracts acoustic evidence associated with one physical property. For a property $a$, we write

$$f_a(x) = s_a(x), \qquad s_a(x) \in [0, 1], \tag{66}$$

where a larger value indicates stronger measured evidence for the corresponding property. The evaluators are designed around interpretable signal-level cues and are shared by both M301 and M302.

### D.1 Decay

The decay evaluator measures whether an acoustic excitation is followed by a clear and sustained reduction in energy. Impact peaks are first detected from a smoothed energy envelope. Nearby peaks corresponding to the same strike are merged, after which a decay curve is fitted independently to the segment following each retained peak.

For detected event $i$, the event-level score is

$$d_i = \left(1 - \exp\left[-\frac{T_{20,i}}{0.45}\right]\right) \mathrm{clip}\left(0.7R_i^2 + 0.3\frac{\mathrm{clip}(D_i, 0, 25)}{25}, 0, 1\right), \tag{67}$$

where $T_{20,i}$ is the estimated time required for a $20\,\mathrm{dB}$ reduction in acoustic energy, $R_i^2$ denotes the goodness of fit of the log-energy linear regression, and $D_i$ is the observed dynamic range in decibels.

Each detected event is weighted by its fitted duration multiplied by $R_i^2$. To prevent short or spurious peaks from diluting a prominent decay, the final score combines the strongest event with the quality-weighted average:

$$s_{\mathrm{decay}} = 0.65 \max_i d_i + 0.35 \frac{\sum_i w_i d_i}{\sum_i w_i}. \tag{68}$$

### D.2 DISTANCE

The distance evaluator combines three complementary acoustic cues: active-frame level, high-frequency attenuation, and late energy.

The level cue uses the 90-th percentile short-time RMS among frames whose energy lies no more than 25 dB below the peak. Let $L_{\mathrm{act}}$ denote this level in dBFS. To characterize spectral darkening, we define

$$H = 10 \log_{10} \frac{E_{200\text{–}4000\,\mathrm{Hz}} + \epsilon}{E_{4000\text{–}8000\,\mathrm{Hz}} + \epsilon}. \tag{69}$$

The three components are combined as

$$\begin{aligned} s_L &= \sigma\left(\frac{-L_{\mathrm{act}} - 16}{5}\right), \\ s_H &= \sigma\left(\frac{H - 12}{10}\right), \\ s_{\mathrm{distance}} &= 0.70 s_L + 0.20 s_H + 0.10 s_{\mathrm{late}}, \end{aligned} \tag{70}$$

where

$$\sigma(z) = \frac{1}{1 + \exp(-z)}. \tag{71}$$

The late-energy term is activated only when a reliable decay fit is available, defined by $R^2 > 0.35$. Otherwise,

$$s_{\mathrm{late}} = 0.5, \tag{72}$$

which acts as a neutral value. This gating prevents continuous sources such as engines or sirens from being interpreted through an unreliable impulse-decay estimate.

### D.3 DOPPLER EFFECT

The Doppler evaluator characterizes time-varying frequency displacement from a spectral-centroid trajectory represented in cents. Three motion hypotheses are considered: a monotonically increasing trajectory for an approaching source, a monotonically decreasing trajectory for a receding source, and a rise–fall trajectory for a pass-by event.

For the pass-by hypothesis, the trajectory is fitted separately on the two sides of the estimated point of closest approach so that the opposite frequency shifts before and after passage do not cancel.

For the best-fitting motion hypothesis, the score is

$$s_{\mathrm{Doppler}} = \left(1 - \exp\left[-\frac{\Delta_c}{300}\right]\right)\left(0.25 + 0.75 R_f^2\right)\left(0.50 + 0.50 q_{\mathrm{dir}}\right), \tag{73}$$

where $\Delta_c$ is the fitted frequency excursion in cents, $R_f^2$ measures the linearity of the fitted trajectory, and $q_{\mathrm{dir}}$ measures agreement with the expected motion direction.

### D.4 ECHO

The echo evaluator measures whether an acoustic event is followed by a distinct delayed repetition. Direct waveform autocorrelation is unsuitable because it can respond strongly to pitch periodicity and stationary tonal components. We therefore compute autocorrelation over an onset-strength representation.

Let $r_t$ denote the short-time RMS envelope. We define the positive change in log energy as

$$o_t = \max\left(\log(r_t + \epsilon) - \log(r_{t-1} + \epsilon), 0\right). \tag{74}$$

After subtracting its temporal mean,

$$\widetilde{o}_t = o_t - \frac{1}{N} \sum_{j=1}^{N} o_j, \tag{75}$$

the normalized autocorrelation is

$$\rho(\tau) = \frac{\sum_t \widetilde{o}_t \widetilde{o}_{t+\tau}}{\sum_t \widetilde{o}_t^2 + \epsilon}. \tag{76}$$

Only delays within the interval

$$\tau \in [35\,\mathrm{ms}, \min(1.5\,\mathrm{s}, 0.55T)] \tag{77}$$

are considered, where $T$ is the duration of the analyzed audio segment.

Let $\rho_{\max}$ denote the largest non-negative autocorrelation peak in this range and $p_\rho$ its prominence relative to the surrounding correlation baseline. The echo score is

$$s_{\mathrm{echo}} = \mathrm{clip}\left(\frac{0.65\rho_{\max} + 0.35p_\rho - 0.04}{0.50}, 0, 1\right). \tag{78}$$

The autocorrelation term measures repeated onset structure, whereas the prominence term requires the repetition to form a distinct delayed peak. Their combination reduces responses to diffuse energy fluctuations and stationary periodic signals that do not contain a clearly separated reflection.

### D.5 Occlusion

The occlusion evaluator captures two acoustic effects: overall transmission loss and preferential attenuation of high-frequency content.

Let $R_{\mathrm{LH}}$ denote the low-to-high-band energy ratio in decibels, using a $1.8\,\mathrm{kHz}$ split, and let $L_{\mathrm{RMS}}$ denote the global waveform level in dBFS. We define

$$\begin{aligned} s_{\mathrm{dark}} &= \sigma\left(\frac{R_{\mathrm{LH}} - 3}{4}\right), \\ s_{\mathrm{att}} &= \sigma\left(\frac{-L_{\mathrm{RMS}} - 30}{7}\right), \\ s_{\mathrm{occlusion}} &= 0.72 s_{\mathrm{dark}} + 0.28 s_{\mathrm{att}}. \end{aligned} \tag{79}$$

The larger weight assigned to spectral darkness reflects the ambiguity of absolute level, which can also vary because of differences in source amplitude or waveform normalization.

### D.6 Resonance

The resonance evaluator accounts for two complementary mechanisms: *free resonance*, in which a mode continues to ring after an excitation, and *driven resonance*, in which a stable acoustic mode persists under continuous excitation.

A candidate spectral mode is characterized by its peak prominence $P$, quality factor $Q$, modal energy fraction $F$, frequency stability $S_f$, and post-impact ringing score $S_r$. Peak evidence is first computed as

$$\begin{aligned} s_P &= 1 - \exp(-P/8), \\ s_Q &= (1 - \exp(-Q/8))\left(1 + \exp\left[\frac{Q - 200}{25}\right]\right)^{-1}, \\ s_{\mathrm{peak}} &= s_P\,(0.30 + 0.70 s_Q). \end{aligned} \tag{80}$$

The suppression of very high $Q$ values reduces responses to isolated spectral bins and unrealistically pure synthetic tones. The free and driven resonance paths are

$$\begin{aligned} s_{\mathrm{free}} &= s_{\mathrm{peak}}\,(0.15 + 0.85 S_r), \\ s_{\mathrm{driven}} &= s_{\mathrm{peak}}\,(0.40 + 0.60 S_f)\left(0.25 + 0.75\,\mathrm{clip}\left(\frac{F - 0.04}{0.28}, 0, 1\right)\right), \\ s_{\mathrm{resonance}} &= \max(s_{\mathrm{free}}, s_{\mathrm{driven}}). \end{aligned} \tag{81}$$

### D.7 REVERBERATION

The reverberation evaluator measures persistent acoustic energy following an excitation using three complementary quantities: estimated reverberation time, late-energy ratio, and decay-fit quality.

The corresponding terms are

$$\begin{aligned} s_{\mathrm{RT}} &= 1 - \exp(-T_{60}/0.75), \\ s_{\mathrm{late}} &= \mathrm{clip}\left(\frac{E_{\mathrm{late}}}{0.22}, 0, 1\right), \\ q_{\mathrm{decay}} &= \mathrm{clip}\left(0.6R_{\mathrm{RT}}^2 + 0.4\frac{D}{30}, 0, 1\right), \\ s_{\mathrm{reverb}} &= \left(0.65s_{\mathrm{RT}} + 0.35s_{\mathrm{late}}\right)\left(0.35 + 0.65q_{\mathrm{decay}}\right)\left(0.40 + 0.60q_{\mathrm{act}}\right). \end{aligned} \tag{82}$$

Here, $T_{60}$ denotes the estimated reverberation time, $R_{\mathrm{RT}}^2$ measures the quality of the decay fit, $D$ is the observed decay range in decibels, and $E_{\mathrm{late}}$ measures late acoustic energy. The activity term $q_{\mathrm{act}}$ is obtained by normalizing the peak-to-low-tail energy drop by $18\,\mathrm{dB}$. This factor suppresses false responses from nearly stationary signals that contain persistent energy but do not exhibit an excitation followed by a reverberant decay.

## E EVALUATOR VALIDATION DETAILS

This section provides additional details of the validation protocol summarized in Sec. 4.4. Because the seven evaluators measure heterogeneous acoustic phenomena, a single validation protocol is not appropriate for all properties. We therefore use positive–negative discrimination when a meaningful non-targeted background distribution can be constructed, ordinal validation for distance, and a separate analysis for echo.

### E.1 POSITIVE–NEGATIVE VALIDATION PROTOCOL

The positive collection contains 226 generated clips, each constructed to explicitly express one target physical property: 36 decay clips, 60 distance clips, 40 Doppler clips, 32 occlusion clips, 32 resonance clips, and 26 reverberation clips. Each clip is used as a positive example only for its designated property.

The negative collection contains 170 acoustically diverse clips spanning animals, speech, household activities, appliances, stationary outdoor ambience, and steady electronic or mechanical sounds. These clips are generated without explicitly requesting any of the evaluated physical effects. Both collections contain clips of $5\,\mathrm{s}$ and $10\,\mathrm{s}$ duration.

For decay, Doppler effect, occlusion, resonance, and reverberation, the same negative collection is used as a shared background distribution. Importantly, these negative examples should not be interpreted as clips in which the corresponding physical property is guaranteed to be absent. Many acoustic phenomena can occur naturally even when they are not explicitly requested. For example, an audio clip that is not prompted for reverberation may still contain environmental reflections, and a non-resonance example may still contain narrow-band spectral structure.

The resulting binary task therefore asks whether clips in which a target property is explicitly induced receive systematically stronger evaluator responses than a broad, non-targeted acoustic background distribution. This protocol measures sensitivity to the intended physical effect rather than classification between strictly property-present and property-absent audio.

For property $a$, let $\mu_a^+$ and $\mu_a^-$ denote the mean evaluator responses over the positive and negative collections, respectively. We report

$$\Delta\mu_a = \mu_a^+ - \mu_a^-, \tag{83}$$

together with ROC–AUC.

As reported in Table 2, all five evaluators assessed under this protocol assign higher mean scores to their corresponding positive examples. Decay exhibits the strongest separation, with $\Delta\mu = 0.458$

and AUC $0.916$, while Doppler reaches AUC $0.842$. Occlusion, resonance, and reverberation obtain AUC values between $0.664$ and $0.681$. The more moderate separation observed for these properties is consistent with the fact that related acoustic cues can also occur naturally in the shared background collection.

### E.2 Ordinal Validation of Distance

Distance requires a different validation protocol because every audible event necessarily occurs at some source–receiver distance. A generic background recording therefore does not form a meaningful "non-distance" class.

Applying the positive–negative protocol mechanically gives an AUC of $0.422$. We do not interpret this value as evidence against the evaluator, because the corresponding binary distinction is not physically well defined. Instead, we evaluate whether the distance score preserves the intended ordering among matched *near*, *mid*, and *far* conditions.

The mean evaluator scores are

$$\bar{s}_{\text{near}} = 0.542, \qquad \bar{s}_{\text{mid}} = 0.588, \qquad \bar{s}_{\text{far}} = 0.662. \tag{84}$$

Across ten matched scenes, nine assign a higher mean score to the far condition than to the corresponding near condition. Six of the ten scenes satisfy the complete monotonic ordering

$$s_{\text{near}} < s_{\text{mid}} < s_{\text{far}}. \tag{85}$$

Distance estimation from monaural generated audio remains intrinsically ambiguous. Source amplitude, spectral content, room response, and generation artifacts can partially imitate or obscure propagation-related distance cues. Nevertheless, the aggregate near-to-far increase and the consistency of the far-versus-near comparison indicate that the evaluator captures useful distance-related acoustic variation.

### E.3 Echo Validation and Its Limitations

Echo is treated separately for a different reason. Its evaluator directly measures delayed repetition through onset-strength autocorrelation and peak prominence, as defined in Appendix D.4, providing an interpretable signal-level measurement for each clip.

In principle, the evaluator could also be validated through positive–negative discrimination. However, the currently available positive echo examples are substantially more homogeneous than the validation sets used for the other properties. Many examples share similar isolated transient sources, delay structures, source types, and synthesis characteristics. A binary classifier-style evaluation on this collection could therefore obtain high apparent discrimination by exploiting generator- or template-specific regularities rather than demonstrating general sensitivity to acoustic echo.

For this reason, we retain echo as a formula-based physical measurement but do not report ROC–AUC for the current collection. A stronger validation set would require a substantially more diverse set of real and synthetic echoes spanning different source signals, delay times, reflection strengths, acoustic environments, and generation procedures. It should additionally contain challenging negative examples with superficially similar temporal structure, including periodic impacts, repeated events, reverberant tails, and amplitude-modulated stationary sounds.

## F Fine-Grained Benchmark Results

### F.1 Per-Metric Results

Table 4 reports the complete per-metric results for all evaluated models, complementing the level-wise summary in Table 3.

Table 4: Fine-grained performance of the ten evaluated audio generators across all PhyAudioBench metrics. M101–M110 correspond to Level 1 acoustic and prompt fidelity, M201–M203 to Level 2 temporal structure, and M301–M302 to Level 3 physical consistency.

| Model | L1 | | | | | | | | | | L2 | | | L3 | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M101 | M102 | M103 | M104 | M105 | M106 | M107 | M108 | M109 | M110 | M201 | M202 | M203 | M301 | M302 | |
| AudioLDM 2 | 1.0000 | 1.0000 | 0.8436 | 0.9805 | 0.9347 | 0.7115 | 0.9625 | 0.4643 | 0.9020 | 0.6818 | 0.7001 | 0.3874 | 0.1257 | 0.0945 | 0.2700 | 0.4783 |
| Make-An-Audio 2 | 1.0000 | 1.0000 | 0.8301 | 0.9952 | 0.9558 | 0.8277 | 0.9986 | 0.6587 | 0.9338 | 0.6364 | 0.8270 | 0.7086 | 0.2434 | 0.2243 | 0.3028 | 0.5801 |
| Auffusion | 1.0000 | 1.0000 | 0.2525 | 0.9485 | 0.9017 | 0.8993 | 0.9806 | 0.6667 | 0.9314 | 0.7652 | 0.8876 | 0.7973 | 0.2573 | 0.2265 | 0.3216 | 0.5853 |
| EzAudio | 1.0000 | 1.0000 | 0.5973 | 0.9675 | 0.9251 | 0.8313 | 0.9819 | 0.6052 | 0.9289 | 0.6364 | 0.8236 | 0.6970 | 0.1809 | 0.2000 | 0.3579 | 0.5645 |
| MAGNeT | 1.0000 | 1.0000 | 0.9541 | 0.9103 | 0.9797 | 0.8130 | 0.8833 | 0.6786 | 0.9020 | 0.7955 | 0.8056 | 0.6798 | 0.1872 | 0.2075 | 0.3381 | 0.5740 |
| MMAudio | 1.0000 | 1.0000 | 0.9042 | 0.9953 | 0.9505 | 0.6353 | 0.9375 | 0.3988 | 0.8211 | 0.7576 | 0.6295 | 0.2639 | 0.0912 | 0.0823 | 0.2342 | 0.4422 |
| Stable Audio Open 1.0 | 1.0000 | 1.0000 | 0.4693 | 0.9176 | 0.8880 | 0.7322 | 0.9958 | 0.4802 | 0.8186 | 0.7045 | 0.7278 | 0.4648 | 0.1513 | 0.1852 | 0.2880 | 0.4950 |
| Tango2 | 1.0000 | 1.0000 | 0.9471 | 0.9964 | 0.9898 | 0.8670 | 0.9264 | 0.6409 | 0.9632 | 0.6970 | 0.8578 | 0.8128 | 0.1828 | 0.2120 | 0.3267 | 0.5966 |
| TangoFlux | 1.0000 | 1.0000 | 0.9306 | 0.8956 | 0.8966 | 0.9069 | 0.9917 | 0.6508 | 0.9926 | 0.6136 | 0.8913 | 0.8379 | 0.4255 | 0.2635 | 0.4087 | **0.6474** |
| Resonate | 1.0000 | 1.0000 | 0.3039 | 0.9904 | 0.9812 | 0.8016 | 0.9681 | 0.5714 | 0.8015 | 0.5606 | 0.7940 | 0.6031 | 0.1338 | 0.1451 | 0.3412 | 0.5171 |

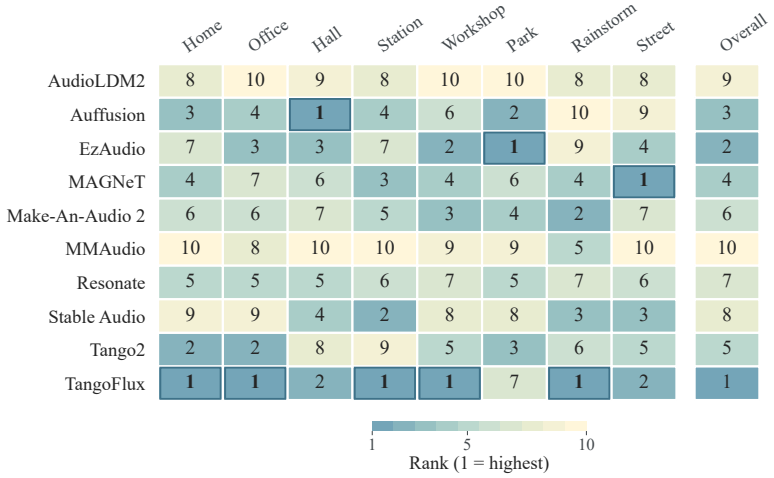



Figure 5: Scene-wise Level 3 rankings across the eight acoustic scenes. Rank 1 denotes the highest Level 3 score within each scene.

### F.2 Scene-Wise Physical Consistency

Figure 5 reports the Level 3 ranking of each model across the eight acoustic scenes. The results show substantial cross-scene variation: four different models rank first in at least one scene, indicating that physical-consistency performance is not uniform across acoustic environments.

## G Model Configurations and Inference Settings

Table 5 summarizes the configurations used to generate all benchmark audio. For each model, the configuration is fixed across benchmark instances, evaluation levels, scenes, and random seeds. All models use a base seed of `20260915` with seed IDs $\{0, 1, 2\}$. The reported results therefore characterize each model under the configuration listed below rather than under an exhaustive model-specific hyperparameter search.

## H Discussion

### H.1 Scalability through Factorized Specification

PhyAudioBench is designed as a structured evaluation framework rather than a fixed collection of manually enumerated prompts. Its scalability comes from separating the semantics of an evaluation target from the individual instances used to instantiate it. Acoustic events, temporal relations, physical properties, and controlled interventions are represented explicitly in the benchmark specification, allowing additional test cases to reuse the same evaluation logic.

Table 5: Configurations of the audio generation models evaluated in PhyAudioBench. Year denotes the year of first public release. All models use a base seed of `20260915` with seed IDs $\{0, 1, 2\}$. DiT, LDM, and NAR denote Diffusion Transformer, Latent Diffusion Model, and Non-Autoregressive Transformer, respectively. SR denotes sample rate. Guidance scales and schedulers use the corresponding default settings.

| Model | Year | #Params | Architecture | Focus | SR | Steps |
|---|---|---|---|---|---|---|
| AudioLDM 2 | 2023 | 350M | LDM | Audio/Music/Speech | 44.1 kHz | 25 |
| Make-An-Audio 2 | 2023 | 937M | LDM | T2A | 44.1 kHz | 25 |
| Auffusion | 2024 | 1.1B | LDM | T2A | 16 kHz | 25 |
| EzAudio | 2024 | 596M | DiT | T2A | 44.1 kHz | 25 |
| MAGNeT | 2024 | 300M | NAR | SFX | 44.1 kHz | $[20, 10, 10, 10]$ |
| MMAudio | 2024 | 157M | Flow Matching | V2A/T2A | 44.1 kHz | 25 |
| Stable Audio Open 1.0 | 2024 | 1.06B | DiT | Audio/Music | 44.1 kHz | 20 |
| Tango 2 | 2024 | 866M | LDM | T2A | 44.1 kHz | 100 |
| TangoFlux | 2024 | 515M | Flow Matching | T2A | 44.1 kHz | 25 |
| Resonate | 2026 | 470M | Flow Matching | T2A | 44.1 kHz | 25 |

This factorization enables the benchmark to scale along several axes without redesigning the evaluation protocol. New acoustic scenes can be introduced by constructing additional event compositions under the same specification schema. Difficulty can be increased by adding sound sources, extending event chains, or associating more physical properties with a scene while keeping individual events semantically identifiable. Likewise, new controlled interventions can be introduced by specifying the affected physical attributes and their expected response directions, after which the existing pairwise evaluation procedure can be reused. The current benchmark therefore represents one instantiation of a broader structured test space rather than an exhaustive enumeration of possible acoustic scenes.

The same separation also makes the evaluation components modular. Property-specific evaluators are decoupled from prompt construction and event routing: a stronger evaluator for an existing physical property can replace the current implementation without changing the benchmark specification or the controlled-response formulation. Extending PhyAudioBench to a new physical property would require an appropriate measurement for that property, but the same event grounding, property routing, and intervention framework can remain unchanged. This modularity is particularly important for physical audio evaluation, where the quality of available acoustic measurements is likely to improve independently of the benchmark construction itself.

### H.2 Limitations and Scope

PhyAudioBench evaluates physical consistency through observable acoustic evidence rather than direct access to the underlying physical process. The proposed physics-aware evaluators are therefore *proxy measurements*: they quantify acoustic signatures associated with distance, Doppler effect, reverberation, decay, resonance, echo, and occlusion, but cannot establish the physical cause of those signatures with certainty. This limitation is fundamental when multiple physical mechanisms produce similar spectral, temporal, or energy patterns. Our evaluator validation measures whether the proposed signals respond systematically to their intended properties, but should not be interpreted as proving a unique physical explanation for every generated waveform.

The benchmark also favors controlled and explicit prompts. This design is intentional because minimal interventions make physical responses easier to attribute and provide cleaner supervision for evaluation. However, naturally occurring user requests may describe physical conditions only implicitly, use more ambiguous language, or leave relevant environmental factors unspecified. Performance on PhyAudioBench therefore characterizes physical consistency under structured and controlled conditions rather than the full ambiguity of unconstrained real-world prompting.

Finally, the current intervention design primarily isolates one targeted physical change at a time. Real acoustic environments often involve coupled physical effects, where changes in geometry, materials, motion, and propagation interact simultaneously. Although the structured specification can associate an intervention with multiple affected attributes, PhyAudioBench does not yet system-

atically evaluate higher-order interactions among multiple simultaneous interventions. Extending controlled-response evaluation from isolated factors to such coupled physical changes is an important direction for future work.