Benchmarking Single-Factor Physical Video-to-Audio Generation

Tingle Li[1,2*] Siddharth Gururani[2*] Kevin J. Shih[2*] Gantavya Bhatt[3] Sang-gil Lee[2]
Zhifeng Kong[2] Arushi Goel[2] Gopala Anumanchipalli[1] Ming-Yu Liu[2]
[1]UC Berkeley [2]NVIDIA [3]University of Washington
https://research.nvidia.com/labs/dir/flatsounds/

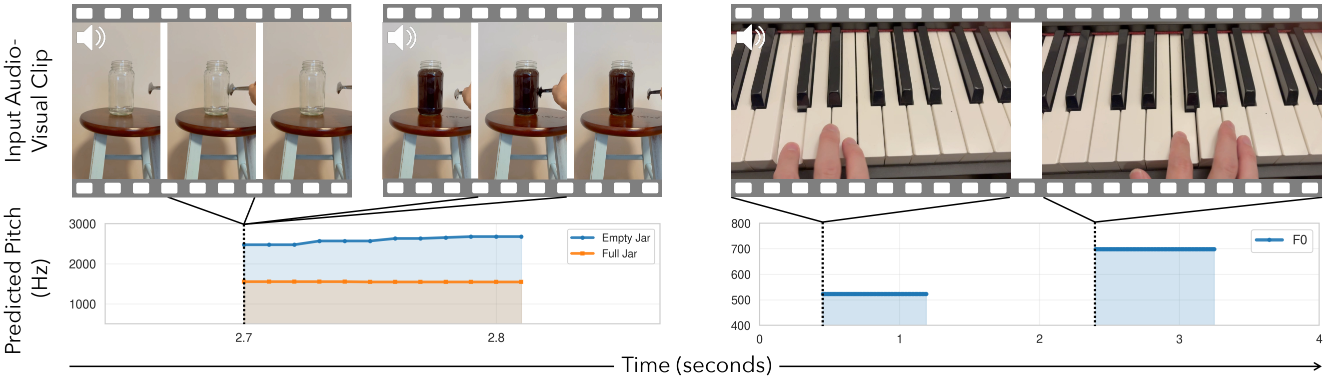

Figure 1. **FlatSounds for video-to-audio physical benchmark.** We test whether current video-to-audio models generate sound that reflects controlled changes in physical factors. For counterfactual pairs, we time-warp videos so that only a single physical factor (e.g., jar fullness) differs while impact timing remains aligned, and compare generated physical features such as pitch (left). For single-video tests, we probe within-clip consistency and directional trends, such as increasing pitch for ascending piano key presses (right).

Abstract

Generative video-to-audio (V2A) models produce highly plausible soundtracks, but it remains unclear whether they capture the underlying physical processes. Existing evaluations emphasize perceptual realism and overlook physical correctness under controlled interventions. In this paper, we introduce FlatSounds, a benchmark that audits the physical reasoning of V2A models through: 1) controlled counterfactual pairs in which a single physical factor is varied, and 2) single-video pattern tests that probe internal consistency and directional trends. These settings test whether the generated audio correctly reflects specific physical properties and timings. Our evaluation of state-of-the-art models reveals a consistent trade-off: models rely more on text captions than the visual stream to infer physics and semantics. Captions generally improve physical and semantic accuracy, but paradoxically degrade temporal alignment. Our results highlight the need to move beyond audio quality toward learning physical processes directly from pixels. Finally, we find that our physics-based metrics correlate strongly with human preference tests on our own data.

1. Introduction

A core aspiration of artificial intelligence is to build general-purpose world models [23, 33], systems that can simulate, reason, and interact with the physical world. A simulation of reality, however, is incomplete if it is silent, as sound is a rich physical signal that reveals latent properties of the world, such as material, fullness, or unseen dynamics, that are often ambiguous to vision alone [3, 32, 61]. The task of video-to-audio (V2A) generation [46, 71] serves as a critical test bed for this grand challenge. To successfully generate audio from video, a model must implicitly simulate the physical processes that produce sound. When a metal spoon strikes a glass (Fig. 1), the resulting acoustic texture [40] is strictly governed by the object's geometry, material composition, and interaction dynamics [16, 44]. A model that truly understands this event is not merely pattern matching, but also simulating an internal physics engine [4, 54].

However, this perspective reveals a stark deficiency in the current evaluation. Recent V2A models [10, 11, 36, 69] produce remarkably plausible sounds, but their success is largely measured by distributional and semantic metrics (e.g., FAD and CLAP) [31, 64] that capture surface-

---

[*] Equal contribution

level *plausibility* rather than deep *physical understanding* [15, 26, 45, 57]. This method obscures a fundamental question: to what degree do these models truly capture the dynamics behind how and why a sound is emitted from physical interaction?

We argue that current benchmarks [11, 49, 60] are insufficient because they test *correlation*, not *causal responsiveness*. Generating a plausible *clink* for a glass does not imply a correct internal model of the underlying physics. We must evaluate the model's response to controlled causal interventions [48], where individual physical factors are systematically manipulated. How does the sound change when the striker's material shifts from metal to wood, or when a container's fullness is modified [8]? A model with a robust physical world model will correctly modulate acoustic properties such as attack time or fundamental frequency [25, 72], while a model built on plausibility alone might fail.

To address this gap, we introduce FlatSounds, a new V2A physical evaluation framework to systematically audit the physical dimension of V2A models. Our benchmark complements conventional metrics by introducing two evaluation modes: controlled counterfactual testing and single-video pattern analysis. We curate time-warped video pairs in which a single physical factor (e.g., material, geometry, environment) is modified while other factors remain fixed, and we design single-video tests for internal consistency (e.g., repeated identical impacts) and directional trends (e.g., ascending pitch). Notably, for the counterfactual pairs, we use time-warping to precisely align impact timings, ensuring the isolated physical variable is the sole cause of any acoustic change. We then use metrics grounded in physics to measure if the change in generated audio correctly reflects the video's delta, and use timing-related metrics to measure temporal alignment.

Our evaluation of state-of-the-art V2A models [11, 36, 52, 69] under varied conditioning settings yields a striking insight: while text captions generally improve semantic alignment and physical correctness, they paradoxically degrade temporal synchronization. Across most models, removing captions actually *improves* certain timing-related metrics. This exposes a critical bottleneck in the video encoder. Current models do not truly see the physics; they preferentially read it from text when available, ignoring the precise visual cues for timing and physical interaction. They effectively "cheat" by relying on explicit text descriptions rather than emergent visual understanding. The pronounced performance drop in physical accuracy when captions are removed confirms this dependency, highlighting a fundamental weakness in current visual representation learning for physical processes. Our key contributions are:

- We introduce a benchmark that reframes V2A evaluation from plausibility to physical correctness, providing a new tool to audit whether models respond causally to controlled changes in physical factors.
- We curate a dataset and protocols with two complementary modes: time-warped factual–counterfactual pairs that isolate a single physical variable, and single-video consistency tests for repeated patterns and directional trends, enabling joint assessment of physical correctness and temporal alignment.
- We show that current V2A models tend to rely on text for physical reasoning at the expense of visual temporal alignment, revealing a core deficiency in video encoders and reframing the central challenge as one of visual physical understanding.

2. Related Works

Video-to-audio generation. The field of V2A generation has evolved from specialized models for specific interactions [46] to open-domain systems. Current methods employ diverse architectures, including auto-regressive transformer models [27, 53, 56], diffusion, flow-matching, and MaskGIT models for high-fidelity output [37, 38, 47, 55, 62, 66], and frameworks for video-audio co-generation [24, 68], which typically involve parallel transformer blocks for the video and audio modality with some form of cross-modality fusion block. In contrast, a dominant and highly successful approach, which includes several of our evaluated baselines [11, 52, 69], involves a two-stage process: first, adapting pre-trained text-to-audio diffusion models [35] with visual adapters [28, 50], or employing multimodal joint training with text-audio data [11]. This reliance on text is even more explicit in models such as ThinkSound [36], which integrates a multi-modal LLM to first generate a textual chain-of-thought reasoning before synthesizing the audio. While these methods achieve state-of-the-art plausibility, their architectural reliance on text motivates our need for a benchmark that can audit whether visual-physical grounding is being ignored. Our contribution is a new benchmark to evaluate this critical dimension.

Video-to-audio evaluation. Current V2A evaluation is primarily focused on plausibility, using datasets and metrics designed to test in-the-wild correlations. The most common evaluation datasets are large-scale, unconstrained collections of internet videos, such as AudioSet [20], VGGSound [9]. These are essential for training, but as uncontrolled collections, they do not provide the paired ground truth needed for causal intervention analysis. Furthermore, creating the necessary controlled interventions post-hoc via video manipulation remains an unsolved challenge [65]. The primary evaluation metrics fall into two main categories. First are distributional and semantic metrics, such as FAD [31], CLAP [64], and ImageBind [22]. Second are temporal synchronization metrics, which typically use specialized models like Synchformer [28] to estimate the weighted temporal offset. Comprehensive benchmark suites, such as AV-Benchmark [11], Movie Gen Audio Bench [49], and Kling-Audio-Eval [60], package these semantic and temporal metrics into robust toolkits for evaluating broad performance. While this landscape is mature, it is designed to measure plausibility and correlation. Our work fills a critical gap by providing a new framework to answer a causal question using controlled interventions.

Physical and counterfactual evaluation. Recent work in evaluating generative models has diverged into two key trends. The first focuses on physical reasoning benchmarks across visual domains. Recognizing the limits of plausibility, frameworks have emerged to audit physical understanding in video generation models. These include VLM-based assessments [5, 41], object tracking metrics [63], human-aligned metrics [26], object-level specificity [2, 43], and real physics experiments [67]. In parallel, for vision-language models, benchmarks like PhysBench [13] and PAI Bench [70] audit an understanding of concepts like mass and density. We extend this line of inquiry from visual properties to the audio domain. The second trend is causal and counterfactual reasoning [48]. This framework is increasingly used to diagnose and correct "shortcut learning" [19]. Benchmarks in visual question answering [1, 30] use counterfactual interventions to force models to learn true causal structures. Our work proposes to synthesize these two threads: we audit acoustic-physical properties using a counterfactual intervention framework.

Acoustics. The link between the physical properties of an object and its sound is well-established [7, 18]. Acoustics is a broad field of physics that provides formal models linking an object's physical attributes, such as geometry, material stiffness, and boundary conditions, to its resulting sound texture [40], including its modal resonances (which determine pitch) and high-frequency damping (which determines timbre) [16, 44]. On the perceptual side, psychoacoustics studies the human perception of these physical-acoustic links [6, 42]. A key concept in this field is the Just Noticeable Difference (JND) [14, 72], which defines the perceptual threshold for changes in features such as fundamental frequency (pitch) or decay rate (duration). Prior work in this area has specifically characterized the perceptual salience of features such as attack time, which strongly correlates with human perception of material hardness (e.g., metal and wood) [17]. Our benchmark leverages this body of work by selecting a suite of objective, perceptually-relevant acoustic measures for evaluating a model's physical understanding, moving beyond correlation-based scores.

3. Benchmark Metrics

We evaluate two complementary aspects of V2A models: 1) temporal alignment, and 2) whether the generated sound responds correctly to single-factor physical changes in the conditioning video.

3.1. Temporal Alignment

We measure audio-video alignment on impact-style events. Our benchmark videos heavily feature actions such as tapping, scratching, plucking, clapping, and smacking. These actions typically produce sounds with short attacks and distinct peaks in their amplitude envelope. We manually annotate the onset times of such events in each video (see Sec. 4.2 for details).

Given the annotated event times, we evaluate the generated audio by detecting onset candidates using an onset-strength detector with an envelope-based fallback, and checking whether at least one detected event lies within an adaptive temporal window around each annotation. This yields an event recall score that we term Hit Coverage (%). We focus on recall rather than precision, since additional unannotated sounds (e.g., ambient noise) may be physically plausible and should not be penalized. We also measure the Timing Error as the average deviation (in ms) of detected hits from the ground truth annotation. Finally, since we generate multiple audio samples per video, we define Perfect Align as the percentage of these generations that achieve 100% Hit Coverage.

3.2. Physical Correctness

Instead of requiring absolute accuracy in pitch ranges or attack times, we focus on directional consistency under controlled physical interventions. When a single relevant factor in the video is changed, we ask whether the corresponding change in the generated audio has the expected direction. For instance, if a clapping video is moved from a heavily furnished living room to a wide stairwell, we expect an increase in reverberation. Our benchmark realizes such single-factor changes across several sound attributes (see visualizations in the supplement) and measures whether the predicted attribute changes in the correct direction.

Temporal envelope features. We capture impact dynamics and material-dependent damping through three envelope-level quantities:

- **Attack Time**: The time it takes a sound to reach its peak amplitude after onset. Replacing a soft material with a harder one (e.g., foam with metal) decreases attack time.
- **Decay Rate**: How quickly a sound decays after the attack. A sheet of metal in firm contact with a table is more strongly damped and should exhibit a higher decay rate than the same sheet freely hanging in the air.
- **Temporal Modulation**: The strength of rhythmic or amplitude fluctuations over time. Shaking a jar of coins should yield stronger temporal modulation than shaking a jar of sand.

Figure 2. **FlatSounds dataset.** How should sound change when we manipulate specific visual properties of a scene? FlatSounds contains indoor recordings of everyday household objects producing sound under controlled variations. Many clips are arranged into time-aligned counterfactual pairs in which we alter a single factor (material, environment, texture, etc.) while keeping the rest of the scene fixed, allowing us to test whether generated audio changes in the expected way. We show the distribution of objects, hits, and metrics (left), as well as example frames of several counterfactual pairs presented in the dataset (right).

Room acoustics. We measure changes in environment using two standard reverberation measures:

- **Room Reverberation Time (RT60)**: The time it takes the sound level to decay by 60dB. The RT60 of a large hall is typically on the order of several seconds, while a small absorptive room has an RT60 below one second.
- **Direct-to-Reverberant Sound Ratio (DRR)**: The ratio between the energy of the direct sound and the energy of the reverberant field. It quantifies the clarity of an environment: a large hall has a lower DRR than a small room since the reverberant energy is relatively stronger.

Spectral and tonal structure. We probe material and excitation changes via spectral and pitch descriptors:

- **Fundamental Frequency (F0)**: The lowest-frequency component of a periodic sound. It is the primary physical correlate of perceived pitch [12]. F0 should ascend or descend as we play ascending or descending notes on any instrument.
- **Spectral Centroid**: The center of gravity of the magnitude spectrum in each STFT frame. The spectral centroid is perceived as the sound's "brightness" or "sharpness" [29, 59]. Harder materials should have an increased spectral centroid [17, 21]. For example, an impact on glass should produce a higher centroid than a similar impact on wood or on a sofa.
- **Spectral Flux**: The average frame-to-frame change in spectral magnitude [34]. Spectral flux is related to the perceived "roughness" or "busyness" of a sound [72]. For instance, smoothly cutting paper with scissors should exhibit lower spectral flux than tearing paper.
- **Spectral Rolloff**: The frequency below which a fixed proportion (typically 85%) of the total spectral energy is contained [34]. Stepping on dry leaves should yield a higher rolloff frequency than stepping on wet leaves, due to the wider spread of energy at higher frequencies.

4. Benchmark Construction

We describe the construction of FlatSounds, the video dataset that supports our alignment and physical correctness benchmarks. The dataset is deliberately compact but tightly controlled: each clip is recorded with a specific physical factor and metric in mind, enabling targeted counterfactual and single-video tests rather than uncontrolled coverage.

4.1. Data Collection

FlatSounds comprises 185 unique indoor clips (Fig. 2), all of which are used in both our single-video alignment analysis and our physical correctness evaluations. Each clip is paired with a human-written caption and timestamps of sound-producing events.

We design the following criteria for collecting video clips: 1) each clip should be between 5 and 10 seconds; 2) the sound-producing event should be roughly centered within the frame; 3) the interaction produces clear, high-energy onsets with a distinct peak in the audio energy envelope; 4) sound events should occur roughly 1-5 times per video, with at least half a second in between each event to facilitate event detection. We focus on impact-style and other clear-onset events because their discrete timing and relatively clean causal structure make single-factor attribution and temporal-alignment auditing more reliable.

All videos are recorded indoors using smartphone cameras and onboard microphones. We focus on everyday physical interactions between household objects (jars, utensils), furniture and structural elements (doors, floors), human hands, and instruments. To probe room acoustics, we record matched interactions in acoustically distinct environments, such as near cushioned sofas *vs*. hard walls, and in narrow corridors or stairwells *vs*. more absorptive living spaces.

The clips are organized around the metrics introduced in Sec. 3. For temporal metrics, we record impact and short-friction events that differ in material, contact geometry, or damping (e.g., the same metal object struck on a table *vs*. hanging freely). For spectral metrics, we design clips that systematically vary pitch, material hardness, or texture (e.g., gravel *vs*. wood impacts). For environmental metrics, we record the same gesture across rooms that differ in size and absorption. Many of these clips are further grouped into factual-counterfactual pairs that isolate a single manipulated factor (material, fullness, environment, or action), while others are used as single-video pattern tests (e.g., monotonic pitch sequences or repeated identical hits). We refer to this benchmark as FlatSounds to emphasize its focus on everyday physics within an apartment-like setting.

4.2. Annotation

Each video is annotated with a text caption and exact timings for sound events of interest. In order to introduce a degree of automation into this process, we narrow our set of sound-generating events to ones that produce a distinct peak when plotting the sound energy envelope. This limits us to common impact-generating sounds such as clapping, tapping, playing musical notes, and fast friction events.

Sound event timing annotation. To obtain precise event timings, we follow a semi-automatic procedure. For each audio track, we compute a short-time Fourier transform and derive a frame-wise energy envelope by taking the RMS of the magnitude spectrum across frequencies. We then apply `scipy.find_peaks` [58] with conservative detection thresholds to get candidate timestamps of onset peaks. In practice, background noise and long-tailed resonances, especially from ringing metal, often introduce multiple false-positive peaks (see the supplementary material for details). We therefore perform a manual verification step: inspect each candidate peak, remove spurious detections, add any missed onsets, and discard clips for which a clean set of event peaks cannot be reliably extracted. The resulting verified timestamps serve as ground truth for our alignment metrics and as anchor points for constructing time-aligned counterfactual pairs.

Creating counterfactual pairs. We construct counterfactual video pairs to evaluate the model's causal responsiveness. Our goal is to temporally align the sound events to the best of our ability, such that the video pairs only differ in the object, action, or background setting that we alter.

We identify pairs from FlatSounds using ground-truth audio annotations and expert knowledge of acoustics, annotating each pair with at least one metric change from Sec. 3.2. At test time, we check that the generated models comply with the direction of the annotated metric change (e.g., we expect an increase in Temporal Modulation from video A to B). The counterfactual video must contain at least as many annotated sound events (hits) as the factual video. We use only the first N hits for alignment if there are more. We record multiple versions of each interaction with varying hit counts to enable flexible pairing.

For temporal alignment, we warp the counterfactual video to match the factual video's event timings. We set time-annotated sound events as anchor points, then stretch or compress the time between anchor points such that the annotated peaks occur at the target timestamps. Frames between anchor points are resampled to achieve the target frame count and segment duration. The time-warping procedure may result in unnatural motion if timestamps differ substantially, but we find results generally good without introducing visual artifacts. For single sound events, warping is unnecessary as we can achieve alignment through frame-trimming alone.

5. Experiments

Here, we evaluate recent V2A models using our benchmark, as well as existing metrics such as FAD, KL, Inception Score, ImageBind, CLAP, DeSync. We aim to demonstrate that our FlatSounds benchmark and physics-informed metrics evaluate V2A model quality in an interpretable and effective manner, and would serve as a good complement to existing evaluation frameworks.

5.1. Experimental Setup

Evaluated models. We evaluate a representative set of state-of-the-art V2A models: FoleyCrafter [69], Hunyuan-V2A [52], MMAudio [11], ThinkSound [36]. To test the hypothesis that models rely on explicit textual cues for physics, we also include MMAudio-Phys, our own fine-tuned variant of MMAudio using physics-aware captions gathered using custom prompts with Omni-captioner [39]. We also evaluate the contribution of the caption during inference by evaluating all models with and without captions.

Datasets and evaluation protocols. We evaluate the available models on both VGGSound [9] and FlatSounds. On VGGSound, we employ all standard evaluation protocols: Fréchet Audio Distance (FAD, computed with PANN,

Table 1. Per-metric average confidence on FlatSounds-Physics. "Temporal Mod." denotes Temporal Modulation, and "Avg." reports the average confidence over all metrics. All metrics are higher-is-better (↑).

Method	Attack Time	Decay Rate	F0	Spectral Centroid	Spectral Flux	Spectral Rolloff	Temporal Mod.	RT60	DRR	Avg.
MMAudio-Phys (w/ Caption)	0.290	0.310	**0.334**	**0.368**	0.321	**0.395**	0.237	0.310	0.189	**0.306**
Hunyuan-V2A (w/ Caption)	0.267	**0.320**	0.326	0.332	**0.403**	0.305	0.283	0.247	0.262	0.305
Hunyuan-V2A (w/o Caption)	0.232	0.306	0.258	0.294	0.383	0.300	0.320	0.271	**0.300**	0.296
MMAudio-Phys (w/o Caption)	**0.293**	0.279	0.310	0.318	0.309	0.347	0.263	**0.316**	0.164	0.289
ThinkSound (w/ Caption)	0.243	0.188	0.198	0.267	0.174	0.294	0.299	0.232	0.157	0.228
MMAudio (w/ Caption)	0.209	0.175	0.228	0.259	0.253	0.223	0.266	0.232	0.189	0.226
MMAudio (w/o Caption)	0.200	0.171	0.230	0.265	0.237	0.243	0.196	0.217	0.230	0.221
ThinkSound (w/o Caption)	0.213	0.194	0.249	0.266	0.246	0.248	0.225	0.177	0.152	0.219
FoleyCrafter (w/o Caption)	0.180	0.179	0.182	0.241	0.294	0.215	0.299	0.176	0.188	0.217
FoleyCrafter (w/ Caption)	0.158	0.157	0.177	0.234	0.267	0.230	**0.333**	0.143	0.147	0.205

Table 2. Overall performance on FlatSounds-Physics. All metrics are higher-is-better (↑).

Method	Confidence	Hit Coverage	Perfect Align	CLAP
MMAudio-Phys (w/ Caption)	**0.306**	82.65	59.82	0.630
Hunyuan-V2A (w/ Caption)	0.305	90.21	69.31	0.633
Hunyuan-V2A (w/o Caption)	0.296	**91.50**	**70.50**	0.593
MMAudio-Phys (w/o Caption)	0.289	83.69	61.00	0.602
ThinkSound (w/ Caption)	0.228	74.81	51.52	0.573
MMAudio (w/ Caption)	0.226	75.02	52.03	**0.642**
MMAudio (w/o Caption)	0.221	75.81	51.50	0.616
ThinkSound (w/o Caption)	0.219	75.05	51.10	0.548
FoleyCrafter (w/o Caption)	0.217	83.39	60.70	0.548
FoleyCrafter (w/ Caption)	0.205	66.52	44.70	0.573

PaSST, and VGG embeddings) [31], KL, Inception Score (IS) [51], ImageBind Score (IB) [22], CLAP score [64], and DeSync [28]. On FlatSounds, we run all aforementioned standard evaluations in addition to our proposed metrics for alignment and physical correctness, as described in Sec. 3.

5.2. FlatSounds Metric Implementation

Temporal alignment metrics. Our temporal alignment metrics (Sec. 3.1) evaluate whether the generated audio from a given video produces energy peaks at the annotated timestamps extracted from the ground-truth audio. These metrics are computed on all 185 captioned and timing-annotated clips (FlatSounds-Single).

Physical correctness metrics. Our physical correctness metrics operate on a set constructed from 185 unique indoor clips referred to as FlatSounds-Physics. Using time-warped pair construction, we get 178 paired video tests containing at least one annotated metric with an expected direction of change ($|\Delta|$), and 90 single-video tests for physical metrics that do not require counterfactual pairs (internal consistency and directional trends within a video). This gives us a total of 268 test cases for FlatSounds-Physics.

It would be meaningless to judge directional changes in metrics if the generated audio is too far from the expected content to even compare. We alleviate this with a "soft gate" to ensure audio is both temporally aligned and semantically plausible *before* judging its physical correctness. The final Confidence score for any metric is computed from per-seed weighted votes. For each seed, we define a quality weight that equally balances temporal alignment and semantic plausibility; for pair comparisons, the temporal and semantic terms each take the minimum of the factual and counterfactual scores for that seed. Audio that is poorly synchronized or semantically incorrect is not discarded but has its influence on the final score proportionally reduced. Finally, we report Confidence for each metric as the weighted proportion of seeds that satisfy the expected physical trend. We provide a detailed formulation in the supplement.

Statistical testing. Our benchmark's Confidence score is derived from statistical voting over 10 seeds.

- **Increase/Decrease**: For pair comparison, a seed only votes if its delta $|\Delta|$ exceeds a robust effect-size threshold, $\tau = \max(2\% \text{ of mean}, 25\% \text{ of robust\_std})$. Seeds with sub-threshold changes ($|\Delta| \leq \tau$) or NaN values (e.g., no hit detected) are counted as failures.
- **Ascending/Descending**: For single-video monotonicity tests (e.g., piano scale), we extract per-hit F0 sequences in $\log_2$ space. For $n = 2$ hits, we check the sign of the difference. For $n \geq 3$, we use Spearman's rank correlation ρ with adaptive thresholds ($|\rho| \geq 0.40$ for $n \leq 4$, 0.30 for $5 \leq n \leq 7$, and 0.25 for longer sequences). Seeds with insufficient valid hits ($n < 2$) are counted as failures.
- **No Change**: For pair comparison, we use a strict equivalence test: the 95% CI of the mean delta must fall entirely within a wider band $[-\tau_{eq}, +\tau_{eq}]$. For single-video tests, we compute a robust coefficient of variation against a per-metric threshold based on JND [72].

5.3. Physical Benchmark Performance

We first audit models using FlatSounds-Physics, which isolates causal reasoning by ensuring identical event timings.

Physical correctness evaluation. The primary finding from FlatSounds-Physics is that all current models struggle with physical reasoning. As summarized in Table 2, the physical Confidence score remains low across models, indicating that current models struggle to learn acoustic physics from pixels. Per-metric analysis (Table 1) shows MMAudio-Phys (w/ Caption) highest on average, indicating that physics-aware captions provide a noticeable boost.

Table 3. Comparison of V2A models on VGGSound and standard benchmark metrics. All models are evaluated under two conditions: with (w/) and without (w/o) text captions. ↓ indicates lower is better. ↑ indicates higher is better. Best results are in bold.

Method	**FAD-PASST ↓**	**FAD-PANN ↓**	**FAD-VGG ↓**	**KL-PANNS ↓**	**KL-PASST ↓**	**IS ↑**	**IB ↑**	**DeSync ↓**
MMAudio-Phys (w/ Caption)	54.73	**3.97**	**0.61**	**1.37**	**1.16**	**18.49**	**34.89**	0.405
MMAudio-Phys (w/o Caption)	66.31	4.38	0.64	1.75	1.49	13.43	33.11	0.399
MMAudio (w/ Caption)	65.86	4.89	1.08	1.68	1.40	17.59	33.03	0.445
MMAudio (w/o Caption)	63.84	4.50	0.80	2.12	1.82	14.45	32.46	0.436
FoleyCrafter (w/ Caption)	182.46	18.45	3.06	2.25	2.24	15.31	25.87	1.195
FoleyCrafter (w/o Caption)	191.30	19.53	3.82	2.50	2.42	10.89	28.27	1.172
Hunyuan-V2A (w/ Caption)	78.38	10.02	2.11	2.04	1.79	15.29	31.66	0.340
Hunyuan-V2A (w/o Caption)	114.51	15.06	2.12	2.52	2.28	7.59	31.19	**0.326**
ThinkSound (w/ Caption)	**52.44**	4.82	0.71	1.44	1.26	17.98	29.24	0.455
ThinkSound (w/o Caption)	62.32	5.02	0.75	1.64	2.18	10.30	26.74	0.433

Table 4. Temporal comparison on FlatSounds-Single, where Hit Coverage (%) and Timing Error (ms) are reported.

Method	**Hit Coverage ↑**	**Timing Error ↓**
Ground Truth	97.12 ± 1.72	17.25 ± 2.64
Hunyuan-V2A (w/o Caption)	**68.55 ± 3.52**	**44.34 ± 1.04**
Hunyuan-V2A (w/ Caption)	65.21 ± 3.81	44.76 ± 1.01
MMAudio-Phys (w/o Caption)	56.46 ± 2.77	46.63 ± 1.05
MMAudio-Phys (w/ Caption)	50.69 ± 4.23	51.34 ± 1.09
FoleyCrafter (w/o Caption)	49.74 ± 4.25	49.32 ± 1.09
FoleyCrafter (w/ Caption)	48.85 ± 3.07	51.48 ± 1.12
ThinkSound (w/o Caption)	36.34 ± 3.58	53.15 ± 1.19
ThinkSound (w/ Caption)	33.74 ± 3.61	53.66 ± 1.21
MMAudio (w/o Caption)	31.95 ± 3.88	56.20 ± 1.17
MMAudio (w/ Caption)	31.12 ± 3.85	57.67 ± 1.20

Most models show higher Confidence with captions, while FoleyCrafter is the only exception, with captions slightly reducing Confidence.

Difficulty of physical properties. Our per-metric analysis reveals a consistent difficulty ordering. On average across models, spectral properties such as Spectral Flux, Spectral Centroid, and Spectral Rolloff are the easiest metrics, while DRR is the most challenging, with Decay Rate and Attack Time also remaining difficult. This suggests that current models capture frequency-domain characteristics more readily than fine-grained temporal dynamics and some acoustic environment cues.

5.4. Semantic Plausibility Evaluation

Next, we evaluate all models on semantic plausibility metrics from the standard benchmark. These results, gathered from both VGGSound and the FlatSounds-Single set of videos, establish a core trade-off inherent in current architectures.

As shown in Table 3, adding captions typically improves semantic plausibility. MMAudio-Phys (w/ Caption) attains the best results on most semantic metrics, with ThinkSound (w/ Caption) achieving the best FAD-PASST. This trend holds among most models: providing text often lowers FAD and KL scores, while raising IS and IB scores. This is mirrored in our FlatSounds-Single (Table 5), where captions generally reduce FAD and KL, and also improve IB, while IS does not improve consistently. This result is intuitive: explicit text guidance aids the model in generating semantically correct and high-fidelity audio.

5.5. Temporal Alignment Evaluation

We evaluate all models for temporal alignment with respect to videos, using our temporal alignment metrics on FlatSounds videos, and DeSync on both FlatSounds and VGGSound. In all cases, we observe that captions tend to degrade temporal synchronization. On VGGSound (Table 3), Hunyuan-V2A (w/o Caption) yields the best DeSync overall, and for every model DeSync is lower (better) without captions. This pattern is corroborated on our FlatSounds-Single (Table 4), where Hit Coverage is consistently higher without captions. Timing Error is also consistently lower without captions for all models, further confirming that text competes with precise visual timing.

5.6. Human Evaluation

Finally, we conduct a human study on a subset of our benchmark data to validate our metrics. The study uses 40 videos, sampled to broadly cover the main recording settings in our benchmark. We set up the evaluation as a pairwise preference test: each trial presents a randomly selected source video together with two generated audio tracks from two randomly selected models. The raters are asked to select the preferred video-audio pair, taking into account synchronization, presence of audio hallucinations, and overall physical plausibility. Due to the significant cost of pairwise comparisons, we limit our human study to captioned versions of MMAudio-Phys, MMAudio, FoleyCrafter, Hunyuan-V2A, and ThinkSound. The final results of the human study are summarized in Table 7, where we see broad agreement between the human-study ranking and the ranking induced by our benchmark when restricted to the captioned models in Table 2. Using the ELO ranking as an oracle, we calculated the Spearman rank correlation between each metric as com-

Table 6. Spearman rank-correlation of V2A metrics *wrt.* ELO ranking.

Metric	**Value**
FAD-PASST	0.7
FAD-PANN	0.5
FAD-VGG	0.6
KL-PANNS	0.5
KL-PASST	0.3
IS	0.3
IB	0.5
DeSync	0.7
CLAP	0.2
Confidence	**0.9**
Hit Coverage	**0.9**
Perfect Align	**0.9**

Table 5. Comparison across V2A models on FlatSounds-Single, where all metrics are presented with 95% confidence intervals.

Method	**FAD-PASST** $\downarrow$	**FAD-PANN** $\downarrow$	**FAD-VGG** $\downarrow$	**KL-PANNS** $\downarrow$	**KL-PASST** $\downarrow$	**IS** $\uparrow$	**IB** $\uparrow$	**DeSync** $\downarrow$
MMAudio-Phys (w/ Caption)	861.47 ± 10.80	**64.17 ± 1.51**	7.81 ± 0.23	**2.13 ± 0.07**	2.65 ± 0.07	4.60 ± 0.13	0.297 ± 0.005	0.723 ± 0.019
MMAudio-Phys (w/o Caption)	919.67 ± 10.69	73.21 ± 1.56	7.91 ± 0.21	2.48 ± 0.08	2.89 ± 0.07	4.67 ± 0.15	0.287 ± 0.004	0.717 ± 0.019
MMAudio (w/ Caption)	865.95 ± 10.02	66.74 ± 1.59	9.29 ± 0.25	2.30 ± 0.08	2.82 ± 0.07	4.89 ± 0.18	0.313 ± 0.004	0.715 ± 0.018
MMAudio (w/o Caption)	894.99 ± 10.32	73.65 ± 1.68	9.32 ± 0.24	2.57 ± 0.09	2.92 ± 0.07	4.89 ± 0.14	0.301 ± 0.005	0.711 ± 0.017
FoleyCrafter (w/ Caption)	870.77 ± 9.32	103.48 ± 2.31	12.38 ± 0.34	3.18 ± 0.09	2.50 ± 0.07	4.62 ± 0.15	0.277 ± 0.004	1.169 ± 0.021
FoleyCrafter (w/o Caption)	978.88 ± 8.33	102.11 ± 1.75	13.94 ± 0.56	3.65 ± 0.09	3.17 ± 0.08	**5.16 ± 0.21**	0.255 ± 0.003	1.152 ± 0.021
Hunyuan-V2A (w/ Caption)	**833.82 ± 11.33**	76.33 ± 1.67	8.51 ± 0.23	2.44 ± 0.09	**2.40 ± 0.07**	4.91 ± 0.14	**0.315 ± 0.004**	0.683 ± 0.018
Hunyuan-V2A (w/o Caption)	893.03 ± 10.07	80.55 ± 1.67	8.96 ± 0.24	2.87 ± 0.09	2.84 ± 0.07	4.94 ± 0.15	0.299 ± 0.004	**0.679 ± 0.019**
ThinkSound (w/ Caption)	972.98 ± 13.71	77.58 ± 1.45	**6.89 ± 0.22**	2.70 ± 0.08	2.86 ± 0.07	4.93 ± 0.11	0.235 ± 0.004	0.754 ± 0.019
ThinkSound (w/o Caption)	1044.44 ± 11.34	93.80 ± 1.57	8.22 ± 0.27	3.30 ± 0.08	3.22 ± 0.07	4.98 ± 0.16	0.210 ± 0.004	0.760 ± 0.019

Table 7. ELO rating summary of our pairwise human evaluation study. Results are shown in descending order of rating, which largely agrees with the ordering induced by FlatSounds-Physics in Table 2. Full pairwise winrates are provided in the supplementary material.

Model	**ELO Rating** $\uparrow$
Hunyuan-V2A (w/ Caption)	**1556**
MMAudio-Phys (w/ Caption)	1550
ThinkSound (w/ Caption)	1509
MMAudio (w/ Caption)	1447
FoleyCrafter (w/ Caption)	1438

puted on FlatSounds videos, reporting the absolute value (higher is better) in Table 6. We find that our FlatSounds metrics (Confidence, Hit Coverage, Perfect Align) all correlated strongly, with the best-performing standard metrics being FAD-PASST and DeSync. This result suggests that our metrics, as measured on our FlatSounds data, may be more effective than most other standard metrics, while being interpretable and quick to compute. Notably, our alignment metrics correlated more strongly than DeSync. This suggests that FlatSounds may be more effective for identifying alignment issues, though our metrics may not easily extend to arbitrary videos without the ability to cleanly identify sound events.

5.7. Overall Analysis

Our experiments collectively expose a fundamental bottleneck in modern V2A models: the video encoder. The results repeatedly show a central paradox that text captions improve semantic correctness but simultaneously degrade temporal alignment, pointing to a deep-seated architectural flaw. Current models treat text as the primary source of what to generate (the semantic category) while treating video as a secondary source of when to generate (the timing). The degradation in temporal alignment indicates a processing conflict: the model, in prioritizing text, tends to lose track of the precise visual cues for onset.

Conversely, the pronounced drop in semantic quality in the no-caption setting highlights this dependency. Most importantly, the consistently low physical confidence scores reveal that the video encoder, on its own, is not learning to extract physical properties from pixels. It cannot *see* the difference between metal and wood in a way that informs the audio synthesis. The model's limited understanding of physics is not an emergent property of visual simulation but is merely *parroted* from the text prompt when available.

We argue that the current V2A's focus must shift. The challenge is no longer just improving audio synthesis quality. The central problem is building video encoders that can internalize the rich physical and semantic information currently provided as a textual "cheat". Until this is solved, V2A models will remain naive script-readers, not true physical world models capable of reasoning from observation.

6. Conclusion

Our work introduces FlatSounds, a new benchmark that reframes video-to-audio generation as a principled audit of a model's implicit understanding of the physical world. It moves beyond plausibility to audit causal responsiveness using controlled counterfactual interventions and single-video pattern tests. Our experiments reveal a consistent limitation: current models struggle to learn physical processes from pixels, with or without text conditioning. We find they are also deeply dependent on text for semantic understanding, and that this reliance creates a fundamental trade-off: semantic and physical correctness are typically gained at the cost of temporal synchronization. This work provides a tool to systematically measure this gap, reframing the central challenge for the field. The goal is no longer just about improving audio synthesis, but building visual representations that can internalize physical processes from pixels. We will release code, dataset, and models on our project site.

Limitations and broader impacts. Our current benchmark is focused on single-factor interventions within indoor environments. It does not yet capture the complexity of compound physical interactions (e.g., simultaneous changes in force, material, and geometry) or the full spectrum of in-the-wild acoustic phenomena. We believe scaling this causal framework to more complex scenarios is a critical future direction. Furthermore, as models improve on these metrics, they will generate more causally convincing audio, which increases the potential for misuse in creating misleading media. Responsible development and detection strategies will remain essential.

Acknowledgements. We thank Yisi Liu, Yanshi Liu, Xinhui Li, and Mingxuan Cai for guidance and assistance with video recording. We also thank those nearby who put up with our many rounds of noisy recordings while we were collecting FlatSounds.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10044–10054, 2020. 3

[2] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 3

[3] Piyush Bagad, Makarand Tapaswi, Cees GM Snoek, and Andrew Zisserman. The sound of water: Inferring physical properties from pouring liquids. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 1

[4] Renée Baillargeon. The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell Handbook of Childhood Cognitive Development*, pages 47–83, 2002. 1

[5] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[6] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997. 3

[7] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994. 3

[8] Patrick A Cabe and John B Pittenger. Human sensitivity to acoustic information from vessel filling. *Journal of Experimental Psychology: Human Perception and Performance*, 26 (1):313, 2000. 2

[9] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020. 2, 5

[10] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18781, 2025. 1

[11] 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 (CVPR)*, pages 28901–28911, 2025. 1, 2, 3, 5

[12] Alain de Cheveigné. Pitch perception. In *Oxford Handbook of Auditory Science: Hearing*. Oxford University Press, 2010. 4

[13] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025. 3

[14] Fernando del Solar Dorrego and Michelle C Vigeant. A study of the just noticeable difference of early decay time for symphonic halls. *The Journal of the Acoustical Society of America*, 151(1):80–94, 2022. 3

[15] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16890–16902, 2022. 2

[16] Andy Farnell. *Designing sound*. Mit Press, 2010. 1, 3

[17] Daniel J Freed. Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *The Journal of the Acoustical Society of America*, 87(1):311–322, 1990. 3, 4

[18] William W Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993. 3

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 3

[20] Jort F Gemmeke, Daniel PW 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 2

[21] Bruno L. Giordano and Stephen McAdams. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, 119(2):1171–1181, 2006. 4

[22] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 2, 6

[23] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018. 1

[24] Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Alper Canberk, Kwot Sin Lee, Vicente Ordonez, and Sergey Tulyakov. Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19373–19385, 2025. 2

[25] David Howard and Jamie Angus. *Acoustics and psychoacoustics*. Routledge, 2013. 2

[26] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2024. 2, 3

[27] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *The British Machine Vision Conference (BMVC)*, 2021. 2

[28] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329, 2024. 2, 6

[29] Paul Iverson and Carol L. Krumhansl. Isolating the dynamic attributes of musical timbre. *The Journal of the Acoustical Society of America*, 94(5):2595–2603, 1993. 4

[30] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017. 3

[31] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Frechet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019. 1, 2, 6

[32] Andrew J Kunkler-Peck and Michael T Turvey. Hearing shape. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1):279, 2000. 1

[33] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *OpenReview*, 62(1):1–62, 2022. 1

[34] Alexander Lerch. *An Introduction to Audio Content Analysis: Music Information Retrieval Tasks and Applications*. Wiley-IEEE Press, Hoboken, N.J, 2 edition, 2023. 4

[35] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning (ICML)*, 2023. 2

[36] Huadai Liu, Jialei Wang, Kaicheng Luo, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 1, 2, 5

[37] Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see-video to audio generation through text. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 101337–101366, 2024. 2

[38] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 48855–48876, 2023. 2

[39] Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations (ICLR)*, 2026. 5

[40] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011. 1, 3

[41] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning (ICML)*, 2025. 3

[42] Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012. 3

[43] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. 3

[44] James F O'Brien, Chen Shen, and Christine M Gatchalian. Synthesizing sounds from rigid-body simulations. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 175–181, 2002. 1, 3

[45] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14277–14286, 2023. 2

[46] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2413, 2016. 1, 2

[47] Santiago Pascual, Chunghsin Yeh, Ioannis Tsiamas, and Joan Serrà. Masked generative video-to-audio transformers with enhanced synchronicity. In *European Conference on Computer Vision (ECCV)*, pages 247–264. Springer, 2024. 2

[48] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017. 2, 3

[49] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2, 3

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PmLR, 2021. 2

[51] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 6

[52] Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025. 2, 5

[53] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5, 2023. 2

[54] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11 (1-2):13–29, 2005. 1

[55] Kun Su, Xiulong Liu, and Eli Shlizerman. From vision to audio and beyond: A unified model for audio-visual representation and generation. In *International Conference on Machine Learning (ICML)*, 2024. 2

[56] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 2

[57] Ashvala Vinay and Alexander Lerch. Evaluating generative audio systems and their metrics. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 858–865, 2022. 2

[58] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 5

[59] Gottfried von Bismarck. Sharpness as an attribute of the timbre of steady sounds. *Acustica*, 30:159–172, 1974. 4

[60] Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, et al. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*, 2025. 2, 3

[61] Renhao Wang, Haoran Geng, Tingle Li, Philipp Wu, Feishi Wang, Gopala Anumanchipalli, Trevor Darrell, Boyi Li, Pieter Abbeel, Jitendra Malik, et al. The sound of simulation: Learning multimodal sim-to-real robot policies with generative audio. In *Proceedings of the 9th Conference on Robot Learning (CoRL)*, 2025. 1

[62] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 128118–128138, 2024. 2

[63] Zihan Wang, Songlin Li, Lingyan Hao, Xinyu Hu, and Bowen Song. What you see is what matters: A novel visual and physics-based metric for evaluating video generation quality. *arXiv preprint arXiv:2411.13609*, 2024. 3

[64] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 1, 2, 6

[65] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024. 2

[66] Manjie Xu, Chenxing Li, Xinyi Tu, Yong Ren, Rilin Chen, Yu Gu, Wei Liang, and Dong Yu. Video-to-audio generation with hidden alignment. *arXiv preprint arXiv:2407.07464*, 2024. 2

[67] Chenyu Zhang, Daniil Cherniavskii, Andrii Zadaianchuk, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025. 3

[68] Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions, 2025. 2

[69] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024. 1, 2, 5

[70] Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025. 3

[71] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3550–3558, 2018. 1

[72] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013. 2, 3, 4, 6