PAVAS: Physics-Aware Video-to-Audio Synthesis

Oh Hyun-Bin[1†] Yuhta Takida[2] Toshimitsu Uesaka[2] Tae-Hyun Oh[4] Yuki Mitsufuji[2,3]

[1]POSTECH [2]Sony AI [3]Sony Group Corporation [4]KAIST

Abstract

Recent advances in Video-to-Audio (V2A) generation have achieved impressive perceptual quality and temporal synchronization, yet most models remain appearance-driven, capturing visual-acoustic correlations without considering the physical factors that shape real-world sounds. We present Physics-Aware Video-to-Audio Synthesis (PAVAS), a method that incorporates physical reasoning into a latent diffusion-based V2A generation through the Physics-Driven Audio Adapter (Phy-Adapter). The adapter receives object-level physical parameters estimated by the Physical Parameter Estimator (PPE), which uses a Vision-Language Model (VLM) to infer the moving-object mass and a segmentation-based dynamic 3D reconstruction module to recover its motion trajectory for velocity computation. These physical cues enable the model to synthesize sounds that reflect underlying physical factors. To assess physical realism, we curate VGG-Impact, a benchmark focusing on object–object interactions, and introduce Audio-Physics Correlation Coefficient (APCC), an evaluation metric that measures consistency between physical and auditory attributes. Comprehensive experiments show that PAVAS produces physically plausible and perceptually coherent audio, outperforming existing V2A models in both quantitative and qualitative evaluations. Visit https://physics-aware-video-to-audio-synthesis.github.io.

1. Introduction

Humans effortlessly infer physical properties of the world from both what they see and what they hear [43], and visual information influences humans' auditory perception due to the audiovisual integration property [13] and prediction mechanism [53]. For example, when a hammer strikes metal or a ball bounces on a floor, we intuitively expect the resulting sound to reflect underlying physical factors such as object velocity, mass, and material.

Agnostic about this, recent Video-to-Audio (V2A) generation models, including autoregressive-[39, 57, 69], GAN-[20], and diffusion-based [23, 40, 46, 68, 70, 72, 75, 81] approaches, learn to generate audio that aligns temporally and semantically with video content and have achieved impressive perceptual quality and synchronization, especially with the emergence of latent diffusion frameworks [11, 35, 51]. Yet, they remain appearance-driven: a model may correctly associate a hammering motion with a "metallic clang," but it may fail to modulate the loudness or spectral sharpness according to the strength and dynamics of the impact, producing physically implausible audio. (see Fig. 1-[Top]).

We refer to this discrepancy as a lack of *physical grounding*—implicit modeling between visual dynamics and acoustic behavior. We define physics-aware audio generation as the synthesis of sounds that not only align with visual events but also produce sounds whose acoustic properties vary consistently with measurable physical values, such as object mass and velocities in a video. Our work bridges this gap

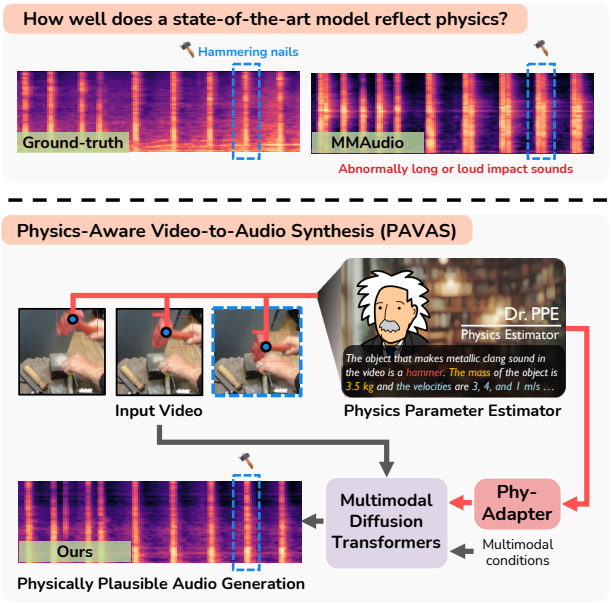

Figure 1. **Physics-Aware Video-to-Audio Synthesis (PAVAS).** **[Top]** Current V2A models often generate physically inconsistent audio. **[Bottom]** We estimate physics values (object-level mass and velocity) from an input video using Physics Parameter Estimator, which are explicitly integrated into a latent diffusion-based model using Phy-Adapter to generate a physically plausible audio.

[†]Work done during an internship at Sony AI.

by explicitly incorporating object-level physical cues into a V2A generation, allowing the model to synthesize sounds that are *perceptually coherent* and *physically consistent*.

To this end, we propose Physics-Aware Video-to-Audio Synthesis (PAVAS), which integrates explicit physical reasoning into a latent diffusion-based generation process (see Fig. 1-[Bottom]). PAVAS includes two key modules: (i) a Physics Parameter Estimator (PPE) that extracts object-level physical quantities from videos, and (ii) a Physics-Driven Audio Adapter (Phy-Adapter), a modulation module that injects the estimated physical parameters into the diffusion model to guide sound synthesis. The PPE consists of a Mass Estimator, which leverages a Vision-Language Model (VLM) [1] to infer object mass from visual and semantic context, and a Velocity Estimator, which combines a text-grounded segmentation model [50] and a dynamic 3D reconstruction model [71] to recover object-level motion and estimate velocity. Surprisingly, we find that our estimators achieve physics value estimation performance comparable to, or even surpassing, that of specialized expert models [6, 80], in both the mass and velocity estimations. Through these components, PAVAS allows the generation process to reflect the object dynamics of real-world interactions, capturing how object motion and mass influence the resulting sound.

Furthermore, we identify the need for an evaluation protocol that explicitly measures physical realism in video-to-audio generation. Existing benchmarks such as VGGSound [7] mainly assess perceptual or semantic alignment, and fail to capture whether generated sounds are consistent with the underlying physical dynamics. To address this, we curate a VGG-Impact benchmark, focusing on object–object interaction events (*e.g.*, collisions, impacts, or bouncing), where physical cues play a critical role. We also present the Audio-Physics Correlation Coefficient (APCC), which quantifies the correlation between estimated physical values (*i.e.*, kinetic energy) and generated audio's attributes (*i.e.*, spectral energy), providing an interpretable measure of physical grounding beyond existing perceptual metrics.

Extensive experiments on VGGSound [7] and VGG-Impact demonstrate that our approach substantially improves physical plausibility while maintaining high perceptual quality. Our model outperforms existing V2A baselines [9, 23, 39, 68–70, 72, 75, 81] on both VGGSound and human evaluations, generating high-quality sounds that are semantically and temporally aligned with video content. Further analysis on VGG-Impact shows that, unlike prior appearance-driven models that exhibit weak or inconsistent correlations on APCC, PAVAS captures physically meaningful relationships between object motion and sound, achieving the closest APCC to ground-truth data. Our main contributions are summarized as follows:

- We introduce PAVAS, a Physics-Aware Video-to-Audio Synthesis pipeline that injects object-level physical parameters—estimated by our reliable physics estimators—into a latent diffusion model via the proposed Phy-Adapter.
- We curate VGG-Impact, a novel benchmark for object–object interaction sounds, and propose APCC, a metric designed to evaluate the physical consistency between visual dynamics and a generated audio.
- We conduct comprehensive analyses of existing V2A models on both VGGSound and VGG-Impact, revealing that prior work often produce physically inconsistent sounds, whereas our method achieves more physically consistent and perceptually realistic audio generation.

2. Related Work

Video-to-audio synthesis Video-to-audio (V2A) generation aims to synthesize realistic sounds that are temporally synchronized and semantically aligned with video content. Recent approaches [9, 20, 27, 40, 46, 57, 68, 72, 75, 81, 82] have significantly advanced this task using increasingly expressive generative models. Autoregressive methods [20, 57] model sequential dependencies effectively, while diffusion-based [17, 58] approaches further improve perceptual quality and synchronization [17, 40, 46, 58, 68, 72, 75, 81]. More recently, MMAudio [9] leverages large-scale text-audio data together with limited video data, establishing a strong foundation for modern V2A synthesis.

Despite these advances, existing audio-visual cross generation methods remain appearance-driven, relying on visual–acoustic correlations [5, 55, 56, 63, 64, 64, 66] without modeling the underlying physical dynamics of real-world interactions. Auxiliary conditioning signals such as onsets, motion energy, or mel cues [9, 23, 68] help synchronization, but these cues often produce smoothed features that fail to capture fine-grained visual dynamics. We instead move beyond perceptual alignment toward physics-aware sound generation, incorporating explicit object-level physical parameters into a V2A generation process.

Physics parameter estimation from video Inferring physical properties from visual data has long been a core goal in visual physics reasoning. Early work explore predicting object dynamics and material attributes such as mass, friction, or elasticity from visual cues or simulated interactions [2, 12, 73, 79]. In this work, we focus on mass and velocity, two quantities that directly determine the magnitude and temporal evolution of physical events in videos. While appearance already encodes material-related cues, mass and velocity govern how objects move, collide, and produce sound, enabling physically aware audio generation.

Recent advances in visual mass estimation, such as NeRF2Physics [80], combine neural radiance fields [24, 25, 29, 42] with vision-language features [48] to infer physical properties from multi-view images. However, these methods require static, calibrated views and cannot operate on dynamic videos. In contrast, we leverage

physics knowledge embedded in Vision-Language Models (VLMs) [1, 19, 65, 77, 78] to estimate object mass from a monocular video, achieving comparable performance while maintaining open-world generalization. Estimating object velocity has been studied less extensively. I-MOVE [54] segments independently moving regions to infer per-object velocity, while most other methods rely on active sensors such as radar or LiDAR [16] are limited to vehicle-centric domains [26, 59]. Building on recent progress in open-vocabulary segmentation [32, 38] and dynamic 3D reconstruction [71], our method estimates reliable physical parameters from unconstrained videos, enabling physics-aware conditioning for video-to-audio synthesis.

Injecting physics cues in video-to-audio generation Recent efforts in Video-to-Audio (V2A) generation have begun to incorporate physical-condition signals to enhance realism. Su *et al*. [61] introduce a diffusion-based model conditioned on physics priors to synthesize impact sounds from silent videos; however, their approach is restricted to generating a drumstick sound and therefore does not generalize to diverse object–object interactions. Saad *et al*. [52] generate acoustic profiles for indoor scenes by controlling material parameters, and SonifyAR [62] proposes an AR sound-authoring pipeline that uses contextual cues such as surface material and interaction type to produce spatialized effects. While these works incorporate contextual or material cues, they are largely designed for indoor or AR-focused settings and do not explicitly model physical factors such as object-level mass or motion dynamics.

3. Physics-Aware Video-to-Audio Synthesis

We propose **Physics-Aware Video-to-Audio Synthesis (PAVAS)**, a novel approach for generating audio that is consistent not only with the visual context but also with the underlying physical dynamics present in a video. PAVAS is based on a latent diffusion architecture and integrates explicit object-level physical reasoning to ensure the generated sounds are physically plausible.

We first describe the diffusion-based backbone and its multimodal conditioning interface (Sec. 3.1). We then provide an overview of the complete pipeline (Sec. 3.2), followed by detailed descriptions of the two key modules: the *Physics Parameter Estimator* (PPE; Sec. 3.3), which extracts object-level mass and velocity from a video, and the *Physics-Driven Audio Adapter* (Phy-Adapter; Sec. 3.4), which incorporates these physical cues into the diffusion model.

3.1. Preliminary

We formulate video-to-audio generation as a latent modeling problem. First, we prepare a variational autoencoder (VAE) [30] pretrained on the mel-spectrogram domain, as well as a pretrained vocoder [36] that converts mel-spectrograms into waveform audio signals. By combining the VAE decoder with the vocoder, compressed audio latent representations can be transformed into waveform signals. Within this framework, a diffusion-based model, denoted as f_θ, is trained to generate audio latent representations conditioned on $\mathbf{Y}$, which typically consists of video frames (and text) in standard video-to-audio generation scenarios.

We employ a flow matching framework to train the latent diffusion model. Let $\mathbf{x}_1$ denote the latent variable. The model is trained to approximate the conditional flow using the following objective:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,q(\mathbf{x}_0),q(\mathbf{x}_1,\mathbf{c})} \left\| f_\theta(t, \mathbf{Y}, \mathbf{x}_t) - u(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) \right\|^2, \tag{1}$$

where $t \in [0, 1]$, $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, and $u(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0$ defines the target flow velocity between the Gaussian source and the data manifold. Minimizing Eq. (1) encourages f_θ to smoothly transport samples from the prior toward realistic audio latents under the given condition $\mathbf{Y}$. The diffusion model enables the generation of $\mathbf{x}_1 \in \mathbb{R}^d$ from randomly sampled Gaussian noise $\mathbf{x}_0 \in \mathbb{R}^d$, by utilizing the learned time-dependent velocity vector field $f_\theta(t, \mathbf{Y}, \mathbf{x}_t)$.

Our approach for incorporating physical cues. We define $\mathbf{Y}$ as a set of multiple modalities, encompassing not only video and text but also physical parameters. During the latent generation process, the latents evolve from $\mathbf{x}_0$ to $\mathbf{x}_1$ through stacked diffusion transformer (DiT) [44] blocks that propagate cross-modal information from $\mathbf{Y}$. In our approach, all conditional modalities—visual, textual, and physical—are projected into a unified hidden space through lightweight encoders and are synchronized across diffusion timesteps using positional embeddings. This backbone provides a stable generative trajectory and serves as the foundation for introducing our physics-aware conditioning.

3.2. Overview

Figure 2 illustrates the overall pipeline of PAVAS. This leverages a multimodal diffusion transformer architecture [9, 11, 35], consisting of transformer blocks shared across visual, textual, and auditory modalities, followed by unimodal blocks specialized for audio decoding.

The *Physics Parameter Estimator* first detects all perceptually moving objects in the input video and estimates object-level, time-invariant mass and per-frame velocities using a combination of vision–language models, segmentation, and dynamic 3D reconstruction. Simultaneously, a vision encoder extracts patch-wise feature embeddings, which are converted into object-centric visual features using masks obtained during velocity estimation. These features provide spatial and semantic context aligned with instance masks, supporting temporal object tracking and facilitating the injection of physical parameters.

Subsequently, the *Physics-Driven Audio Adapter* fuses the physical parameters with object-centric visual features

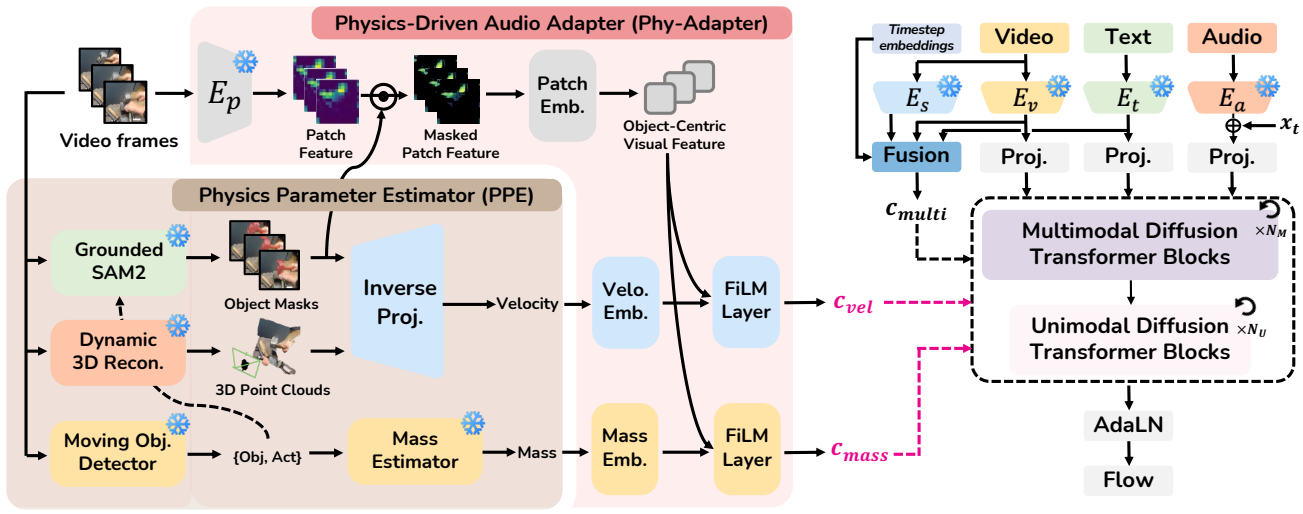

Figure 2. **Overall pipeline of the proposed Physics-Aware Video-to-Audio Synthesis (PAVAS).** Given an input video, the Physics Parameter Estimator (PPE) extracts object-level mass and velocity. These physics cues are encoded by the Physics-Driven Audio Adapter (Phy-Adapter) and injected into the latent diffusion model alongside multimodal conditions. E_p stands for CLIP vision encoder for patch embeddings, E_s for SyncFormer [21] vision encoder, E_v for CLIP vision encoder for flattened visual tokens, E_t for CLIP text encoder, and E_a for VAE/STFT-based audio encoder. Dashed lines indicate conditioning pathways, and magenta highlights physics-based conditioning.

to form temporally aligned, physics-aware representations. To ensure stable and effective fusion, we introduce a Δ-*modulation* mechanism, which enables gradual injection of the fused features into the multimodal transformer blocks of the diffusion backbone. This process guides the model to generate audio that is both perceptually realistic and physically grounded in the underlying object dynamics.

3.3. Physics Parameter Estimator

The physics parameter estimator quantifies object-wise *mass* and *velocity* from unconstrained videos, providing a quantitative link between visual dynamics and physically grounded sound generation. Given an input video $\{I_\ell\}_{\ell=1}^{L}$, the estimator identifies all moving objects $\mathcal{O} = \{o_i\}_{i=1,2,\dots}$ and computes, for each object o_i, a time-invariant mass m_i (in kilograms) and a time-varying velocity sequence $\{v_i^\ell\}_{\ell=1}^{L-1}$ (in meters per second) in metric 3D space:

$$\mathcal{P} = \left\{ (m_i, \{v_i^\ell\}_{\ell=1}^{L-1}) \,\middle|\, o_i \in \mathcal{O} \right\}. \tag{2}$$

These quantities are estimated through a unified three-stage pipeline: (i) *Moving-object detection* uses a Vision–Language Model (VLM) [1] to localize dynamic entities and generate textual object-level descriptors; (ii) the same VLM performs *mass estimation* based on visual appearance and textual cues; (iii) a combination of text-grounded segmentation [50] and dynamic 3D reconstruction [71] recovers object-centric geometry and metric-scale velocity.

Moving-object discovery. We first identify entities exhibiting genuine motion, excluding apparent displacement from camera movement. A vision-language model [1] is prompted with instructions to define "moving objects" (see supplementary material for details). The model outputs a structured set $\mathcal{S} = \{s_i\}_{i=1}^{N}$, where each entry s_i consists of a localized moving object o_i (*e.g.*, "runner in striped shirt") and its action a_i (*e.g.*, "sprinting"), *i.e.*, $s_i = (o_i, a_i)$. This text-level representation enables open-world generalization and serves as a semantic interface for mass and velocity estimation.

Mass estimation. For each object o_i in $\mathcal{S}$, physical mass is estimated using the Vision-Language Model (VLM) [1]. Given the object name o_i and action a_i, a textual prompt $\mathcal{T}_{\text{mass}}$ is constructed to instruct the model to infer the object's mass (see supplementary material for details). The prompt and video context are provided to the mass estimator f_{mass}, yielding $m_i = f_{\text{mass}}(I_{1:L}, \mathcal{T}_{\text{mass}})$. Unlike geometry-based methods [60, 80] that require multi-view supervision, our approach operates directly on monocular dynamic videos and generalizes across diverse object categories and scenes. The resulting $\{m_i\}_i$ values are time-invariant and later paired with velocity sequences.

Velocity estimation. Reliable velocity estimation for physics-aware conditioning requires metric-scale, temporally coherent per-object trajectories across arbitrary categories. However, existing monocular depth estimation models [3, 8, 18, 47, 76] do not meet these requirements: they either lack open-vocabulary, instance-specific object handling or fail to provide temporally consistent metric geometry, and therefore cannot yield physically interpretable object velocities. To satisfy these requirements, we combine open-vocabulary segmentation with dynamic 3D reconstruction, enabling object-wise centroid trajectories in metric world coordinates for subsequent velocity computation.

Textual object descriptors from $\mathcal{S}$ are used as prompts for

Florence-2 [74], which outputs bounding boxes aligned to target classes. SAM-2 [49] refines these boxes into pixel-accurate binary masks and propagates them across time, yielding a sequence of object instance masks $\{\mathbf{M}_i^\ell\}_{\ell=1}^{L}$, where $\mathbf{M}_i^\ell \in \{0,1\}^{H\times W}$. CUT3R [71] reconstructs dense 3D geometry for each frame ℓ, producing a set of 3D points (point cloud) $\mathbf{P}^\ell = \{\mathbf{p}^{\ell,k} \in \mathbb{R}^3\}_{k=1}^{K_\ell}$ and corresponding camera extrinsics $(\mathbf{R}^\ell, \mathbf{T}^\ell)$ under a shared world coordinate system, where K_ℓ is the number of point clouds in frame ℓ.

Each object's 3D extent is obtained by inverse-projecting its 2D mask $\mathbf{M}_i^\ell$ onto the reconstructed 3D scene. Each pixel (h, w) in $\mathbf{M}_i^\ell$ is mapped to its corresponding 3D point $\mathbf{p}^{\ell,k}$ via the 2D-3D mapping g_{inv} from CUT3R, aggregating matched points into object-wise 3D points: $\mathbf{X}_i^\ell = g_{\text{inv}}(\mathbf{M}_i^\ell, \mathbf{P}^\ell) = \{\mathbf{p}^{\ell,k} \in \mathbf{P}^\ell \mid \mathbf{M}_i^\ell(h,w) = 1\}$, where k, h, and w range over $1 \le k \le K_\ell$, $1 \le h \le H$, and $1 \le w \le W$. The centroid $\mathbf{c}_i^\ell$ is then computed to specify the spatial position in metric 3D space for each object: $\mathbf{c}_i^\ell = \frac{1}{|\mathbf{X}_i^\ell|}\sum_{\mathbf{x}\in\mathbf{X}_i^\ell}\mathbf{x}$.

Given the video frame rate FPS and $\Delta\tau = 1/\text{FPS}$, displacement and instantaneous velocity are computed as

$$d_i^\ell = \|\mathbf{c}_i^{\ell+1} - \mathbf{c}_i^\ell\|_2, \qquad v_i^\ell = d_i^\ell/\Delta\tau. \tag{3}$$

The resulting velocity sequence $\mathbf{v}_i$ provides a metric-scale trajectory that captures object-level temporal motion for subsequent physical conditioning.

3.4. Physics-Driven Audio Adapter

Physics-Driven Audio Adapter (Phy-Adapter) bridges the physical values estimated in Sec. 3.3 with the multimodal latent diffusion backbone, transforming per-object mass and velocity into temporally aligned conditioning signals. It receives three inputs at each frame ℓ: (i) visual patch embedding $\mathbf{V}^\ell$ from a CLIP-ViT [22] encoder, (ii) binary mask $\mathbf{M}_i^\ell$ for each object from Florence-2 [74] and SAM2 [49], and (iii) physical values $\{m_i, v_i\}$ estimated from our Physics Parameter Estimator. From these, Phy-Adapter constructs physics conditions that are aligned with the visual feature sequence and include all the moving object information. Then, we inject them into the diffusion transformer's conditioning stream. This integration guides the diffusion trajectory toward physically consistent audio generation.

Object feature extraction. The first stage of Phy-Adapter converts raw visual patch embeddings into object-conditioned representations that serve as the base for subsequent physical modulation. For each frame ℓ and detected object o_i, the module localizes the object region within the visual feature map $\mathbf{V}^\ell \in \mathbb{R}^{H\times W\times D_p}$ using its binary segmentation mask $\mathbf{M}_i^\ell \in \{0,1\}^{H\times W}$ obtained from Florence-2 [74] and SAM2 [49]. An object-specific feature per frame is computed via masked summation:

$$\mathbf{f}_i^\ell = \sum_{h,w} \mathbf{M}_i^\ell[h,w] \cdot \mathbf{V}^\ell[h,w,:], \tag{4}$$

where each patch embedding $\mathbf{V}^\ell[h,w,:] \in \mathbb{R}^{D_p}$ encodes local appearance and motion cues. The aggregated vector $\mathbf{f}_i^\ell$ is subsequently projected into a hidden space via an affine linear transformation, followed by the application of a lightweight LayerNorm, yielding $\mathbf{h}_i^\ell \in \mathbb{R}^{D_h}$. Frames where the object is absent or occluded are replaced by a learnable *object-occlusion token* $\mathbf{z}_{\text{obj-occ}} \in \mathbb{R}^{D_h}$, ensuring temporal continuity and stability in the object-level conditioning stream. The resulting set $\{\mathbf{h}_i^\ell\}_{i,\ell}$ provides object-centric spatio-temporal context upon which mass and velocity modulation are subsequently applied.

Mass and velocity modulation. To condition object-centric visual features on physical properties, both mass and velocity are processed through a similar pipeline. Scalar mass m_i is normalized via $\log(1+m_i)$ and z-score normalization with dataset statistics $(\mu_{\text{mass}}, \sigma_{\text{mass}})$, while velocity v_i^ℓ is z-normalized using $(\mu_{\text{vel}}, \sigma_{\text{vel}})$. Each normalized value is expanded using *Fourier feature mapping* [67]:

$$\mathbf{f}_{\text{mass},i} = [\sin(2\pi\omega_k m_i), \cos(2\pi\omega_k m_i)]_{k=1}^{K} \in \mathbb{R}^{2K}, \tag{5}$$

$$\mathbf{f}_{\text{vel},i}^\ell = [\sin(2\pi\omega_k v_i^\ell), \cos(2\pi\omega_k v_i^\ell)]_{k=1}^{K} \in \mathbb{R}^{2K}. \tag{6}$$

These features are projected by separate MLPs into embeddings $\mathbf{e}_{\text{mass},i}$ and $\mathbf{e}_{\text{vel},i}^\ell$ in $\mathbb{R}^{D_h}$. Subsequently, we broadcast $\mathbf{e}_{\text{mass},i}$ and flatten $\mathbf{e}_{\text{vel},i}^\ell$ to reshape them into $\tilde{\mathbf{e}}_{\text{mass},i} \in \mathbb{R}^{LD_h}$ and flatten $\tilde{\mathbf{e}}_{\text{vel},i} \in \mathbb{R}^{LD_h}$, respectively. FiLM [45] coefficients are then generated as $(\gamma_{\text{mass},i}, \beta_{\text{mass},i}) = \text{Linear}(\tilde{\mathbf{e}}_{\text{mass},i})$ and $(\gamma_{\text{vel},i}, \beta_{\text{vel},i}) = \text{Linear}(\tilde{\mathbf{e}}_{\text{vel},i})$, and applied to modulate the object-centric visual feature $\mathbf{h}_i$:

$$\mathbf{h}_{\text{mass},i} = (1+\tfrac{1}{2}\tanh(\gamma_{\text{mass},i})) \odot \mathbf{h}_i + \tfrac{1}{2}\tanh(\beta_{\text{mass},i}), \tag{7}$$

$$\mathbf{h}_{\text{vel},i} = (1 + \tfrac{1}{2}\tanh(\gamma_{\text{vel},i})) \odot \mathbf{h}_i + \tfrac{1}{2}\tanh(\beta_{\text{vel},i}). \tag{8}$$

Mass modulation remains constant across time for each object, controlling global loudness and decay, while velocity modulation is frame-dependent, adapting audio features to instantaneous motion. For frames where the object is absent or occluded, a learnable *velocity-occlusion token* $\mathbf{z}_{\text{vel-occ}} \in \mathbb{R}^{D_h}$ ensures temporal consistency in velocity conditioning.

Aggregation and physics-aware conditioning. Finally, the modulated mass and velocity features are aggregated via gated pooling:

$$\mathbf{c}_{\text{mass}} = \frac{\sum_i G_{\text{mass},i}\mathbf{h}_{\text{mass},i}}{\sum_i G_{\text{mass},i}}, \quad \mathbf{c}_{\text{vel}} = \frac{\sum_i G_{\text{vel},i}\mathbf{h}_{\text{vel},i}}{\sum_i G_{\text{vel},i}}, \tag{9}$$

where $G_{\text{mass},i}, G_{\text{vel},i} = \sigma(\text{MLP}(\mathbf{h}_{\text{mass},i})), \sigma(\text{MLP}(\mathbf{h}_{\text{vel},i}))$ and σ denotes a sigmoid function. We omit the frame index ℓ from variables here for brevity.

We construct a multimodal condition $\mathbf{c}_{\text{multi}}$ by fusing visual features from the CLIP visual encoder [22], synchronization features from Synchformer [21], text embeddings from the CLIP text encoder, and timestep embeddings, following MMAudio [9]. This multimodal condition is injected

into each transformer block through Adaptive Layer Normalization (AdaLN) layers, providing a unified global and dynamic audiovisual context.

To incorporate physical cues while preserving multimodal stability, we introduce a Δ-*modulation* mechanism that augments AdaLN coefficients with zero-initialized residual mixers driven by physics conditions. Rather than directly summing physics features with multimodal conditions, each transformer block refines its modulation parameters ω as

$$\tilde{\omega} = \omega(\mathbf{c}_{\text{multi}}) + \alpha_m \, g_m(\mathbf{c}_{\text{mass}}) + \alpha_v \, g_v(\mathbf{c}_{\text{vel}}), \qquad (10)$$

where g_m, g_v are lightweight zero-initialized MLPs and α_m, α_v are learnable gates controlling their magnitude. ω denotes the AdaLN modulation parameters (scale and shift) computed from the multimodal condition, and $\tilde{\omega}$ represents their physics-augmented counterpart used within each transformer block. This residual formulation ensures that physical cues are injected gradually—allowing the model to adapt mass and motion effects without perturbing the multimodal representation—thereby aligning diffusion dynamics with physically consistent audiovisual behavior (see supplementary material for details on the diffusion transformer blocks).

4. Experiments

We first outline the evaluation setup, and then present thorough analyses of the experimental results. Due to the space limitation, we present more implementation details and additional experiments in the supplementary material.

4.1. Experimental Settings

Implementation details. PAVAS is trained in two stages. In the first stage, we train a multimodal latent diffusion backbone [9, 11, 35] for general video-to-audio generation. Then, we integrate the Physics Parameter Estimator (PPE) and the Physics-Driven Audio Adapter (Phy-Adapter) into the backbone to inject explicit mass and velocity cues. The audio, visual, and text encoders are frozen while the diffusion transformer blocks and conditioning pathways are optimized. The backbone is trained for 300k iterations using AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-6}) with gradient clipping and a batch size of 512. The same optimization settings are used in the second stage, except that the number of iterations and the learning rate are reduced to 30K and 1×10^{-5}, respectively. During physics-aware fine-tuning, physics tokens are replaced with their corresponding empty tokens with probability 0.1 to handle cases where motion cues are missing or not detected. PAVAS does not strictly assume perfectly trackable visible objects: learnable occlusion tokens handle missing object and velocity cues, while text conditioning still enables the generation of off-screen sound sources. We train model variants operating at 16kHz and 44.1kHz, which differ only in backbone capacity.

Dataset. We train models using a combination of multimodal and audio–text corpora. For general video-to-audio learning, we draw video–audio pairs from VGGSound [7] and complement them with large-scale audio–text datasets [10, 28, 41]. For audio–text datasets, the missing visual tokens are replaced with learnable tokens, allowing the multimodal backbone to remain compatible across data sources. When training the physics-aware components in the second stage, only the VGGSound dataset is used. Audio clips from all datasets are clipped to 8-second segments. Following common practice [9, 23, 39], we feed the corresponding VGGSound class label as a text input of the model.

To assess physical realism, we additionally curate VGG-Impact, a subset of the VGGSound test split composed of 10 sound classes and 272 impact moments, where object mass and motion directly influence the produced sound, such as *hammering nails* and *basketball bounce*. We first filter clips based on sound classes. We then manually remove ambiguous scenes with poorly defined contact dynamics and retain momentary interactions where impact strength and object inertia are visually identifiable. This enables targeted evaluation of whether generated audio reflects the underlying physical parameters inferred from video.

Metrics. We evaluate generation performance across established dimensions—distributional fidelity, perceptual quality, semantic alignment, and temporal synchronization. Following prior work [9, 20, 72], we report Fréchet Distance (FD) and Kullback–Leibler divergence (KL) to measure feature distribution similarity between generated and ground-truth audio, using PaSST [34], PANNs [33], and VGGish [14]. Audio quality is measured using the Inception Score (IS). Semantic alignment is evaluated via cross-modal cosine similarity in the ImageBind space [15]. Temporal synchronization is quantified using the DeSync metric from Synchformer [21], which estimates audio–video misalignment.

To assess the physical plausibility of generated audio, we introduce the *Audio–Physics Correlation Coefficient (APCC)*, which measures how changes in physical magnitude relate to changes in acoustic onset strength. For each impact event, we first detect the onset in the audio and measure its spectral strength using a standard onset detection function [4]. We then estimate the change of kinetic energy at the onset time using the object's predicted mass and its pre- and post-impact velocities in the corresponding video. This quantity represents the mechanical energy lost at impact, expected to be radiated as an acoustic impulse [31]. APCC measures the correlation between these two sequences, kinetic energy changes and acoustic onset strengths, within each sound class in VGG-Impact. We compute this correlation for both ground-truth and generated audio and compare the two. A lower APCC-Δ indicates that the generated audio more closely matches the real coupling between kinetic

Method	Params	Physics corr.	Distribution matching					Audio quality	Semantic align.	Temporal align.
		APCC-Δ↓	FD_{PaSST}↓	FD_{PANNs}↓	FD_{VGG}↓	KL_{PANNs}↓	KL_{PaSST}↓	IS↑	IB-score↑	DeSync↓
See & Hear [75]*	415M	0.566	219.0	24.58	5.40	2.26	2.30	8.58	33.99	1.204
V-AURA [69]*◇	695M	0.654	218.5	14.80	2.88	2.42	2.07	10.08	27.64	0.654
VATT [39]†	-	0.673	131.9	10.63	2.77	**1.48**	1.41	11.90	25.00	1.195
Frieren [72]†◇	159M	0.662	106.1	11.45	1.34	2.73	2.86	12.25	22.78	0.851
FoleyCrafter [81]*	1.22B	0.588	140.1	16.24	2.51	2.30	2.23	15.68	25.68	1.225
V2A-Mapper [70]†◇	229M	0.671	84.57	8.40	**0.84**	2.69	2.56	12.47	22.58	1.225
TARO [68]*◇	258M	0.758	159.1	10.49	1.57	2.92	2.67	9.62	22.85	1.169
MMAudio-L [9]†	1.03B	0.536	60.60	4.72	0.97	1.65	1.40	17.40	33.22	**0.442**
PAVAS-L (Ours)	1.04B	**0.378**	**47.38**	**3.99**	1.15	1.55	**1.35**	**17.51**	**35.41**	0.446

Table 1. **Quantitative comparison on the VGGSound test set.** Following the standard evaluation protocol [9, 72], parameter counts exclude pretrained encoders (*e.g.*, CLIP), latent audio encoders/decoders, and the modules that are not used in test time (*e.g.*, vocoders). We report only the Large variants of MMAudio and PAVAS; both operate at 44.1kHz, while all other models run at 16kHz. ∗: results reproduced using publicly released code. †: results evaluated from author-provided samples. ◇: models that do not use text input at test time.

energy changes and spectral onset strength, and thus exhibits stronger physics consistency. Detailed definitions and analyses of APCC are provided in the supplementary material.

4.2. Experimental Results and Analyses

We evaluate our framework on both perceptual and physical dimensions, examining (i) how existing video-to-audio generation models behave under physically grounded evaluation, (ii) how PAVAS improves perceptual and physical quality, and (iii) which design factors contribute to its effectiveness.

How well do existing models reflect physics? To analyze whether current Video-to-Audio (V2A) models generate acoustics that follow real physical behavior, we evaluate nine state-of-the-art Video-to-Audio (V2A) models [9, 39, 68–70, 72, 75, 81], including ours, on the VGG-Impact benchmark using the proposed Audio–Physics Correlation Coefficient (APCC) (See Table 1). Across models, we observe APCC-Δ values frequently exceeding 0.5, indicating noticeable gaps with respect to the ground-truth correlation between physical quantities and acoustic responses. PAVAS obtains the lowest APCC-Δ among evaluated methods, suggesting that it more closely matches the physics–audio relationship present in the dataset. Overall, the results indicate that while existing V2A models may capture semantic alignment, their generated audio only partially captures variations in underlying physical magnitudes.

This motivates incorporating explicit object-level mass and velocity as conditioning signals to promote more physically consistent V2A generation. Since such conditioning is only effective when the underlying physical estimates are sufficiently plausible, we assess both components of the Physics Parameter Estimator (PPE). Across these evaluations, both estimators demonstrate favorable performance, supporting their suitability as conditioning inputs; details and results are provided in the supplementary material.

Does PAVAS improve audio generation quality? Table 1 presents quantitative comparisons across distributional, semantic, and synchronization metrics against state-of-the-art Video-to-Audio (V2A) models. PAVAS achieves consistently favorable performance across all measures, suggesting that incorporating explicit physics cues can enhance not only physical plausibility but also audio quality on in-the-wild video data [7]. Spectrogram analysis in Fig. 3 reveals that existing V2A models often fail to reflect the impact dynamics present in the video (*e.g.*, a trampoline bounce or a tuning-fork strike). In contrast, our method generates physically plausible audio whose spectral patterns better reflect the strength and dynamics of the impact, showing a closer match to the ground-truth spectrogram at the moment of interaction. We further conduct a user study on the VGGSound [7] test set, where participants rate four aspects of the generated audio: audio quality, semantic alignment, temporal alignment, and *physical plausibility*. Table 2 shows that PAVAS achieves favorable subjective scores across all criteria.

Method	Audio qual.↑	Semantic align.↑	Temporal align.↑	Physical plau.↑
See & Hear [75]	1.77±0.84	1.71±1.02	1.49±0.79	1.63±0.92
V-AURA [69]	2.68±1.14	2.90±1.30	2.86±1.28	2.58±1.24
VATT [39]	2.27±0.93	2.51±1.20	2.00±0.98	2.22±1.09
V2A-Mapper [70]	2.75±1.09	2.55±1.35	1.92±0.99	2.12±1.17
TARO [68]	2.44±1.07	2.42±1.25	1.93±0.98	2.08±1.09
MMAudio-L [9]	3.98±0.98	4.14±1.01	4.06±1.02	3.90±1.11
PAVAS-L (Ours)	**4.23±0.77**	**4.47±0.71**	**4.45±0.80**	**4.37±0.84**

Table 2. **User study on the VGGSound test set.** 27 participants rate eight generated audios on four aspects: audio quality, semantic alignment, temporal alignment, and physical plausibility. We report the mean and standard deviation of the Likert [37] scale scores (1–5; strongly disagree, disagree, neutral, agree, strongly agree).

What makes improvements on audio quality? To analyze which components of PAVAS are responsible for the observed improvements in audio generation, we conduct ablation studies summarized in Table 3. Simply training the backbone [9] longer on VGGSound [7] results in only marginal changes across evaluation metrics, suggesting that improvements in our full model arise from the physics-aware components rather than additional adaptation to the dataset.

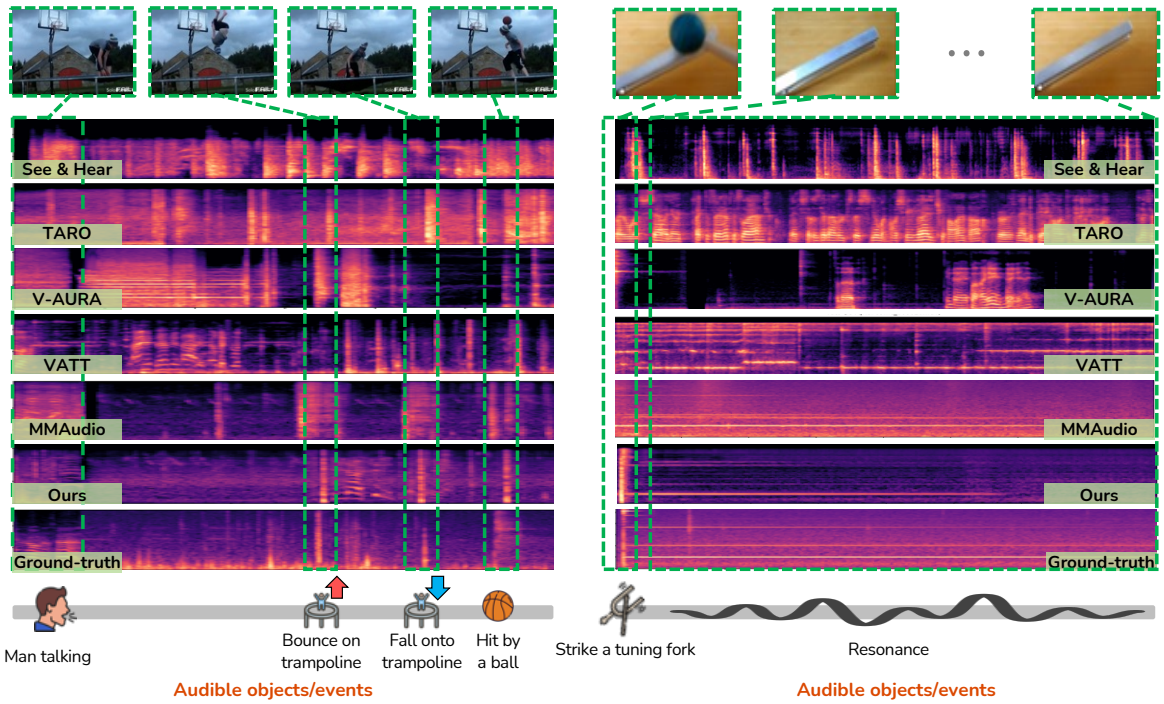

Figure 3. **Qualitative comparison of generated spectrograms.** We visualize spectrograms from existing V2A models [9, 39, 68, 69, 75], our method, and the ground truth. Green dashed lines indicate spectral patterns temporally aligned with visual events in the video, and graphic icons denote audible objects or interactions present in the audio track. PAVAS produces spectral patterns that more closely align with these events, whereas other methods often generate components that are not well aligned with the visual dynamics.

Setting	FD_{PaSST}↓	IS↑	IB-score↑	DeSync↓
(A) Additional Adaptation				
Backbone	**70.19**	**14.44**	29.13	**0.483**
+ Trained longer	71.99	14.34	**29.46**	0.486
(B) Physics Features (w/ Δ-Modulation)				
+ c_{mass} only	66.89	15.94	29.40	0.480
+ c_{vel} only	67.22	15.07	29.33	**0.446**
+ c_{mass} and c_{vel} (ours)	**65.67**	**16.50**	**29.41**	0.448
(C) Injection Strategy (w/ both c_{mass} and c_{vel})				
Direct Summation	67.31	16.30	29.40	0.455
Δ-Modulation (ours)	**65.67**	**16.50**	**29.41**	**0.448**

Table 3. **Ablation study on additional adaptation, physics features, and injection strategies.** (A) examines the effect of training the pretrained backbone longer without physics conditions. (B) investigates the contribution of each physics feature under Δ-modulation. (C) compares summation vs. residual Δ-modulation when both features are used. All models use the S-16kHz backbone and share the same training configuration.

In contrast, introducing physics features leads to consistent improvements: conditioning on mass or velocity individually enhances distributional and perceptual metrics, and combining both yields the best results. We further compare conditioning strategies and find that residual Δ-modulation surpasses direct summation, suggesting that injecting physics cues as residual adaptive modulations within the diffusion transformer allows the model to incorporate physical information more effectively.

5. Conclusion

We presented **PAVAS**, a physics-aware video-to-audio generation method that conditions a latent diffusion model on object-level mass and velocity. Both quantities are estimated from Physics Parameter Estimator (PPE) and injected via Physics-Driven Audio Adapter (Phy-Adapter), enabling the model to produce sounds whose intensity and temporal structure better reflect underlying visual dynamics. To evaluate physical plausibility, an aspect overlooked in prior work, we introduce VGG-Impact and Audio–Physics Correlation Coefficient (APCC), which measure how generated audio follows changes in kinetic energy. Experiments show that mass and velocity conditioning each contribute notable improvements, and their combination yields stronger perceptual quality and better physics consistency than existing models.

Discussion & future direction. PAVAS involves several pretrained vision modules. However, it is not a simple combination of existing components: object-level mass and velocity are inferred without supervised audio–physics labels, and the lightweight Phy-Adapter provides an effective pathway to inject these cues into the diffusion model. Ablations show that this module contributes gains beyond visual features alone, indicating that physical conditioning is meaningful and complementary. Future work may explore more compact adapters, jointly optimized physics estimators, and richer physical factors such as explicit material modeling.

6. Acknowledgments

We sincerely thank Dongseok Shim and Koichi Saito for their insightful internal reviews, and Masato Ishii and Takashi Shibuya for their support in reproducing MMAudio.

O. Hyun-Bin and T.-H. Oh was partially supported by the InnoCORE program of the Ministry of Science and ICT(25-InnoCORE-01, Trust-Enhanced Mutualistic Bio-Embedded AI (30%)), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World (23.3%)), the National Research Foundation of Korea(NRF) funded by the Korea government(MSIT) (No. RS-2024-00451947 (23.3%); No. RS-2024-00358135, Corner Vision: Learning to Look Around the Corner through Multi-modal Signals (23.3%)).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *CVPR*, 2015. 2

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4

[4] Sebastian Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detection. In *the 16th International Conference on Digital Audio Effects (DAFx-13)*, 2013. 6

[5] Lee Chae-Yeon, Oh Hyun-Bin, Han EunGi, Kim Sung-Bin, Suekyeong Nam, and Tae-Hyun Oh. Perceptually accurate 3d talking head generation: New definitions, speech-mesh representation, and evaluation metrics. In *CVPR*, 2025. 2

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 2

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 2, 6, 7

[8] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, 2025. 4

[9] 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 *CVPR*, 2025. 2, 3, 5, 6, 7, 8

[10] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*, 2020. 6

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1, 3, 6

[12] Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. In *ICLR*, 2016. 2

[13] Waka Fujisaki, Naokazu Goda, Isamu Motoyoshi, Hidehiko Komatsu, and Shin'ya Nishida. Audiovisual integration in the human perception of materials. *Journal of vision*, 2014. 1

[14] 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 *ICASSP*, 2017. 6

[15] 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 *CVPR*, 2023. 6

[16] Mian Guo, Kai Zhong, and Xiaozhi Wang. Doppler velocity-based algorithm for clustering and velocity estimation of moving objects. In *2022 7th International Conference on Automation, Control and Robotics Engineering (CACRE)*, 2022. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[18] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, 2025. 4

[19] Nam Hyeon-Woo, Moon Ye-Bin, Wonseok Choi, Lee Hyun, and Tae-Hyun Oh. VLM's Eye Examination: Instruct and Inspect Visual Competency of Vision Language Models. *TMLR*, 2025. 3

[20] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *BMVC*, 2021. 1, 2, 6

[21] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP*, 2024. 4, 5, 6

[22] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021. 5

[23] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *AAAI*, 2025. 1, 2, 6

[24] Kim Jun-Seong, Kim Yu-Ji, Moon Ye-Bin, and Tae-Hyun Oh. Hdr-plenoxels: Self-calibrating high dynamic range radiance fields. In *ECCV*, 2022. 2

[25] Kim Jun-Seong, Mingyu Kim, GeonU Kim, Tae-Hyun Oh, and Jin-Hwa Kim. Factorized Multi-Resolution HashGrid for Efficient Neural Radiance Fields: Execution on Edge-Devices. *IEEE Robotics and Automation Letters*, 2024. 2

[26] Moritz Kampelmühler, Michael G Müller, and Christoph Feichtenhofer. Camera-based vehicle velocity estimation from monocular video. *arXiv preprint arXiv:1802.07094*, 2018. 3

[27] Tornike Karchkhadze, Kuan-Lin Chen, Mojtaba Heydari, Robert Henzel, Alessandro Toso, Mehrez Souden, and Joshua Atkins. Stereofoley: Object-aware stereo audio generation from video. *arXiv preprint arXiv:2509.18272*, 2025. 2

[28] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL*, 2019. 6

[29] GeonU Kim, Kim Youwang, and Tae-Hyun Oh. FPRF: Feed-forward photorealistic style transfer of large-scale 3D neural radiance fields. In *AAAI*, 2024. 2

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Lawrence E Kinsler, Austin R Frey, Alan B Coppens, and James V Sanders. *Fundamentals of acoustics*. John wiley & sons, 2000. 6

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 3

[33] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM TASLP*, 2020. 6

[34] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *INTERSPEECH*, 2022. 6

[35] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 1, 3, 6

[36] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *ICLR*, 2023. 3

[37] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 7

[38] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 3

[39] Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see-video to audio generation through text. In *NeurIPS*, 2024. 1, 2, 6, 7, 8

[40] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023. 1, 2

[41] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM TASLP*, 2024. 6

[42] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[43] Collins Opoku-Baah, Adriana M Schoenhaut, Sarah G Vassall, David A Tovar, Ramnarayan Ramachandran, and Mark T Wallace. Visual influences on auditory behavioral, neural, and perceptual processes: A review. *Journal of the Association for Research in Otolaryngology*, 2021. 1

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 3

[45] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 5

[46] Trung X Pham, Tri Ton, and Chang D Yoo. Mdsgen: Fast and efficient masked diffusion temporal-aware transformers for open-domain sound generation. In *ICLR*, 2025. 1, 2

[47] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 4

[48] 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 *ICML*, 2021. 2

[49] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5

[50] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2, 4

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[52] Mahnoor Fatima Saad and Ziad Al-Halah. How would it sound? material-controlled multimodal acoustic profile generation for indoor scenes. In *ICCV*, 2025. 3

[53] Erich Schröger, Anna Marzecová, and Iria SanMiguel. Attention and prediction in human audition: a lesson from cognitive psychophysiology. *European Journal of Neuroscience*, 2015. 1

[54] Jonathan Schwan, Akshay Dhamija, and Terrance Boult. I-move: Independent moving objects for velocity estimation. In *WACV*, 2020. 3

[55] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound Source Localization is All about Cross-Modal Alignment. In *ICCV*, 2023. 2

[56] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Toward Interactive Sound Source Localization: Better Align Sight and Sound! *IEEE TPAMI*, 2025. 2

[57] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP*, 2023. 1, 2

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[59] Mohammadreza Alipour Sormoli, Mehrdad Dianati, Sajjad Mozaffari, and Roger Woodman. Optical flow based detection and tracking of moving objects for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 3

[60] Trevor Standley, Ozan Sener, Dawn Chen, and Silvio Savarese. image2mass: Estimating the mass of an object from its image. In *Conference on Robot Learning*, 2017. 4

[61] Kun Su, Kaizhi Qian, Eli Shlizerman, Antonio Torralba, and Chuang Gan. Physics-driven diffusion models for impact sound synthesis from videos. In *CVPR*, 2023. 3

[62] Xia Su, Jon E. Froehlich, Eunyee Koh, and Chang Xiao. Sonifyar: Context-aware sound generation in augmented reality. In *UIST*, 2024. 3

[63] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *CVPR*, 2023. 2

[64] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, and Tae-Hyun Oh. Sound2vision: Generating diverse visuals from audio through cross-modal latent alignment. *arXiv preprint arXiv:2412.06209*, 2024. 2

[65] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. AVHBench: A Cross-Modal Hallucination Benchmark for Audio-Visual Large Language Models. In *ICLR*, 2025. 3

[66] Kim Sung-Bin, Kim Jun-Seong, Junseok Ko, Yewon Kim, and Tae-Hyun Oh. Soundbrush: Sound as a brush for visual scene editing. In *AAAI*, 2025. 2

[67] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020. 5

[68] Tri Ton, Ji Woo Hong, and Chang D Yoo. Taro: Timestep-adaptive representation alignment with onset-aware conditioning for synchronized video-to-audio synthesis. In *ICCV*, 2025. 1, 2, 7, 8

[69] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *ICASSP*, 2025. 1, 7, 8

[70] Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *AAAI*, 2024. 1, 2, 7

[71] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2, 3, 4, 5

[72] 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 *NeurIPS*, 2025. 1, 2, 6, 7

[73] Jiajun Wu, Ilker Yildirim, Joseph J. Lim, Bill Freeman, and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *NeurIPS*, 2015. 2

[74] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024. 5

[75] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024. 1, 2, 7, 8

[76] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. In *ICLR*, 2025. 4

[77] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. BEAF: Observing BEfore-AFter Changes to Evaluate Hallucination in Vision-language Models. In *ECCV*, 2024. 3

[78] Moon Ye-Bin, Roy Miles, Tae-Hyun Oh, Ismail Elezi, and Jiankang Deng. RetouchLLM: Training-free Code-based Image Retouching with Vision Language Models. *arXiv preprint arXiv:2510.08054*, 2025. 3

[79] Ilker Yildirim, Jiajun Wu, Yilun Du, and Joshua B. Tenenbaum. Interpreting dynamic scenes by combining a physics engine and bottom-up visual cues. In *Cognitive Science Conference*, 2016. 2

[80] Albert J Zhai, Yuan Shen, Emily Y Chen, Gloria X Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, and Shenlong Wang. Physical property understanding from language-embedded feature fields. In *CVPR*, 2024. 2, 4

[81] 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:2407.01494*, 2024. 1, 2, 7

[82] Lei Zhao, Rujin Chen, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Foleyspace: Vision-aligned binaural spatial audio generation. *arXiv preprint arXiv:2508.12918*, 2025. 2