Physically Grounded Video-to-Audio Generation

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

Abstract

Video-to-audio (V2A) models can now synthesize perceptually plausible and temporally aligned sounds, but they often remain weakly grounded in the physical quantities that shape real acoustic events, such as object mass and motion. We introduce **PAVAS**, a physics-aware V2A framework that estimates object-level mass and velocity from video and injects these cues into a latent diffusion backbone through a lightweight Physics-Driven Audio Adapter. To evaluate this behavior, we curate VGG-Impact, a subset of impact-centric VGGSound clips, and propose the Audio-Physics Correlation Coefficient (APCC), which measures whether generated onset strengths vary consistently with estimated kinetic-energy changes. Across VGGSound and VGG-Impact, PAVAS improves both standard metrics and physics consistency compared with prior strong V2A models.

1. Introduction

Humans infer physical properties of the world from what they see and hear (Opoku-Baah et al., 2021), and visual information influences auditory perception through audiovisual integration and prediction mechanisms (Fujisaki et al., 2014; Schröger et al., 2015). For example, when a hammer strikes metal or a ball bounces on a floor, the sound should reflect factors such as object velocity, mass, and material. Recent video-to-audio (V2A) models have made rapid progress in semantic alignment and synchronization (Iashin & Rahtu, 2021; Xing et al., 2024; Zhang et al., 2024; Wang et al., 2025b; Cheng et al., 2025; Ton et al., 2025), especially with latent diffusion frameworks. Yet most of these systems remain appearance-driven: they may associate a hammering motion with a metallic clang, but fail to modulate loudness or spectral sharpness according to the strength and dynamics of the impact.

We study *physics-aware video-to-audio synthesis*: generating audio that is not only semantically and temporally aligned with a video, but also varies consistently with measurable physical values such as object mass and velocity. The challenge is that in-the-wild videos rarely provide supervised physical labels. Object mass is only indirectly visible, object velocity requires temporally coherent object trajectories, and multiple visible objects may contribute unequally to a generated sound. A useful conditioning mechanism must therefore introduce physical cues while preserving the perceptual quality of a strong V2A backbone.

We propose **PAVAS**, a physics-aware V2A framework that makes physical cues explicit. PAVAS has two components. First, a Physics Parameter Estimator (PPE) estimates object-level mass and velocity from an input video using a vision-language model, open-vocabulary segmentation, and dynamic 3D reconstruction. Second, a Physics-Driven Audio Adapter (Phy-Adapter) converts the estimated values into temporally aligned conditioning signals and injects them into a latent diffusion backbone through residual Δ-modulation. This design keeps the original multimodal representation intact while allowing physical cues to steer the generated audio.

The PPE decomposes the problem into two quantities with direct relevance to object interactions. Mass is estimated from visual and semantic context using a vision-language model, while velocity is recovered from text-grounded segmentation (Ren et al., 2024) and dynamic 3D reconstruction (Wang et al., 2025a). These estimates are then used as conditioning inputs rather than as explicit simulator parameters. This keeps PAVAS compatible with in-the-wild V2A generation, where videos may contain diverse object categories, camera motion, and imperfect observations.

This focus complements prior V2A work that improves perceptual quality or synchronization through stronger generative models and auxiliary cues such as onsets, motion energy, or mel features. These cues are useful, but they do not explicitly model the physical factors that govern object interactions. Mass and velocity are a practical first pair of physical variables for this setting: they can be estimated from unconstrained videos, directly determine kinetic-energy changes at impacts, and provide object-level cues for physically grounded sound generation.

Evaluation is also important. Standard V2A metrics measure distributional fidelity, semantic alignment, or synchro-

[†]Work done during an internship at Sony AI. [1]POSTECH [2]Sony AI [3]KAIST [4]Sony Group Corporation.
Machine Learning for Audio Workshop at ICML 2026.

nization, but they do not directly test whether generated sounds follow the physical dynamics of the video. We therefore curate **VGG-Impact**, a benchmark of impact-centric clips from VGGSound (Chen et al., 2020), and introduce the **Audio-Physics Correlation Coefficient (APCC)**. APCC compares changes in estimated kinetic energy with the strength of generated audio onsets, providing an interpretable measure of physics consistency.

We make three contributions. First, we introduce PAVAS, a V2A model conditioned on object-level mass and velocity estimated from unconstrained videos. Second, we propose VGG-Impact and APCC to evaluate physical grounding in generated audio. Third, we show that explicit physical conditioning improves both standard V2A metrics and physical consistency, with ablations indicating that mass, velocity, and Δ-modulation each contribute to the gains.

2. Physics-Aware Video-to-Audio Synthesis

PAVAS augments a latent diffusion V2A backbone with explicit physical conditions. Given an input video, the model estimates object-level physical parameters, converts them into audio-relevant conditioning tokens, and uses these tokens to modulate the generation trajectory. Figure 1 summarizes the pipeline.

2.1. Backbone and Physical Conditions

We build on a multimodal latent diffusion transformer following recent V2A systems (Esser et al., 2024; Labs et al., 2025; Cheng et al., 2025). Audio is represented in a compressed mel-spectrogram latent space using a VAE and reconstructed with a neural vocoder (Lee et al., 2023). The diffusion model generates an audio latent conditioned on video and text features; in PAVAS, this condition is extended with object-level physical cues. Let $\mathcal{O} = \{o_i\}$ denote moving objects in a video. For each object, PAVAS estimates a time-invariant mass m_i and a frame-wise velocity sequence $\{v_i^\ell\}_{\ell=1}^{L-1}$, forming

$$\mathcal{P} = \{(m_i, \{v_i^\ell\}_{\ell=1}^{L-1}) \mid o_i \in \mathcal{O}\}. \tag{1}$$

These values are not used as hard simulation constraints. Instead, they are projected into the model's conditioning space and guide the diffusion backbone toward audio whose acoustic properties vary with the estimated physical dynamics. In parallel with physical estimation, a vision encoder extracts patch-wise features from the video. Object masks obtained during velocity estimation convert these patch features into object-centric visual features, which provide spatial and semantic context aligned with the estimated physical values. Phy-Adapter then fuses these object-centric features with mass and velocity to produce temporally aligned physics-aware representations.

2.2. Physics Parameter Estimator

The Physics Parameter Estimator (PPE) extracts $\mathcal{P}$ from unconstrained videos without requiring ground-truth physical labels. It has three stages.

Moving-object discovery uses a vision-language model (Bai et al., 2025) to identify entities with genuine motion while excluding apparent displacement due to camera movement. The output is a set of object-action descriptions such as "basketball bouncing" or "hammer striking nail." These text descriptions act as an open-vocabulary semantic interface for mass and velocity estimation.

Mass estimation queries the same vision-language model with the video context and object-action descriptions to infer an approximate mass for each object. Unlike geometry-based methods that require multi-view supervision (Standley et al., 2017; Zhai et al., 2024), this module operates on monocular dynamic videos and leverages visual and semantic context for open-world object categories. The resulting mass values are time-invariant and later paired with velocity sequences.

Velocity estimation combines open-vocabulary segmentation and dynamic 3D reconstruction. Textual object descriptions are converted into object masks using Florence-2 (Xiao et al., 2024) and SAM-2 (Ravi et al., 2024). CUT3R (Wang et al., 2025a) then reconstructs dense 3D geometry for each frame in a shared metric coordinate system. For each object and frame, we inverse-project the object mask onto the reconstructed geometry, compute an object centroid $\mathbf{c}_i^\ell$, and estimate instantaneous velocity as

$$v_i^\ell = \|\mathbf{c}_i^{\ell+1} - \mathbf{c}_i^\ell\|_2 \cdot \mathrm{FPS}. \tag{2}$$

This produces object-wise metric motion trajectories for subsequent physical conditioning.

2.3. Physics-Driven Audio Adapter

The Physics-Driven Audio Adapter (Phy-Adapter) transforms PPE outputs into conditioning signals for the diffusion backbone. For each object, we first extract object-centric visual features by pooling CLIP patch embeddings (Ilharco et al., 2021) within the corresponding segmentation mask. Missing observations are replaced by learned occlusion tokens, allowing the model to handle temporary occlusion or failure of the estimator.

Mass and velocity are then normalized and encoded with Fourier features (Tancik et al., 2020). The resulting embeddings modulate object-centric features using lightweight FiLM-style transformations (Perez et al., 2018). Mass is broadcast across time, capturing global factors such as expected loudness and decay, while velocity remains frame-dependent, adapting audio features to instantaneous motion. For frames where an object is absent or occluded, learned

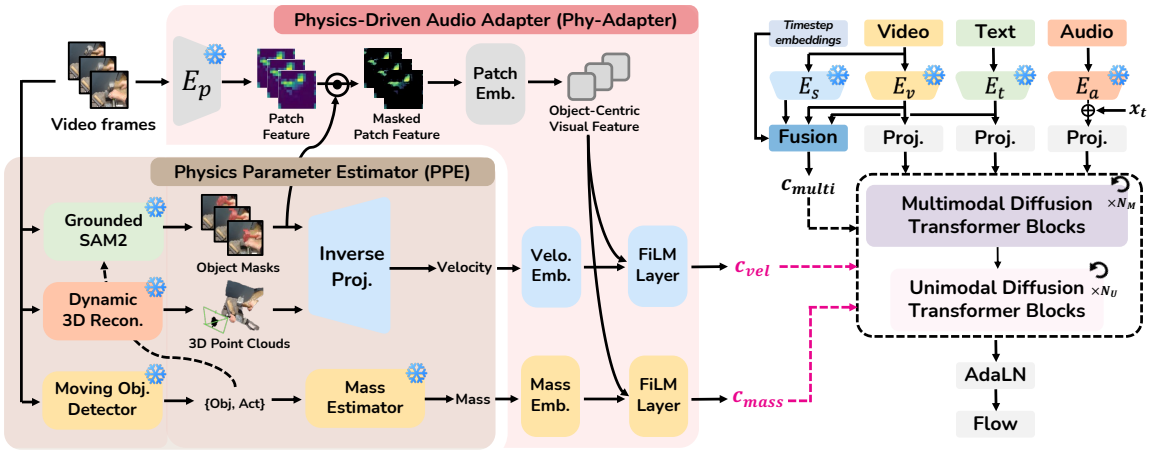

Figure 1. **PAVAS overview.** A Physics Parameter Estimator extracts moving objects, estimates mass, and recovers metric velocity. The Physics-Driven Audio Adapter fuses these values with object-centric visual features and injects them into a latent diffusion V2A backbone via residual Δ-modulation.

occlusion tokens maintain temporal continuity. Object-wise conditioned features are aggregated through gated pooling to form mass and velocity conditions, $\mathbf{c}_{\text{mass}}$ and $\mathbf{c}_{\text{vel}}$.

To incorporate physical cues while preserving multimodal stability, we inject physics through residual Δ-modulation. For each transformer block, the original multimodal condition produces AdaLN parameters $\omega(\mathbf{c}_{\text{multi}})$; PAVAS augments them by zero-initialized residual mixers:

$$\tilde{\omega} = \omega(\mathbf{c}_{\text{multi}}) + \alpha_m g_m(\mathbf{c}_{\text{mass}}) + \alpha_v g_v(\mathbf{c}_{\text{vel}}), \quad (3)$$

where g_m and g_v are small MLPs and α_m, α_v are learnable gates. Because the residual branches are initialized near zero, the model begins from the behavior of the original backbone and gradually learns how mass and velocity should alter the audio generation trajectory. This residual formulation lets the model incorporate mass and motion cues without directly overwriting the multimodal condition.

3. Experiments

3.1. Setup

Training. PAVAS is trained in two stages. We first train the latent diffusion V2A backbone on VGGSound (Chen et al., 2020) together with audio-text corpora (Kim et al., 2019; Drossos et al., 2020; Mei et al., 2024). We then fine-tune with PPE outputs and Phy-Adapter on VGGSound. The audio, visual, and text encoders remain frozen; the diffusion transformer and conditioning paths are optimized. The backbone is trained for 300K iterations with AdamW, learning rate 1×10^{-4}, weight decay 1×10^{-6}, gradient clipping, and batch size 512. The second stage uses the same optimization setup except for 30K iterations and learning rate 1×10^{-5}. During physics-aware fine-tuning, we randomly replace physics tokens with learned empty tokens with probability 0.1, making the model robust to missing or unreliable physical observations.

Training data. For general V2A learning, we combine video-audio pairs from VGGSound with large-scale audio-text datasets. For audio-text data, missing visual tokens are replaced with learnable tokens so that the multimodal backbone remains compatible across data sources. During the physics-aware second stage, we use only VGGSound. Audio clips are clipped to 8-second segments, and following common V2A practice (Jeong et al., 2025; Liu et al., 2024; Cheng et al., 2025), the corresponding VGGSound class label is used as text input.

Benchmarks. We evaluate general V2A quality on VGGSound. To test physical grounding, we curate **VGG-Impact**, a subset of VGGSound test clips containing visible object-object interactions where mass and motion should influence audio, such as basketball bounces, hammer strikes, and other short impact events. It contains 10 impact-centric sound classes and 272 annotated impact moments. We filter by sound class and manually remove ambiguous clips with unclear contact dynamics or invisible impact moments. This enables targeted evaluation of whether generated audio reflects physical parameters inferred from video.

Metrics. For standard V2A quality, we report distributional fidelity using Fréchet Distance and Kullback-Leibler divergence with PaSST, PANNs, and VGGish features (Koutini et al., 2022; Kong et al., 2020; Gemmeke et al., 2017). Audio quality is measured by Inception Score, semantic alignment by ImageBind similarity (Girdhar et al., 2023), and synchronization by DeSync from Synchformer (Iashin et al., 2024). For physical grounding, we use the **Audio-Physics Correlation Coefficient (APCC)**. At each impact event j, we estimate a kinetic-energy change:

$$\Delta E_j = \frac{1}{2} m_j \big((v_j^-)^2 - (v_j^+)^2\big), \quad (4)$$

Table 1. **Quantitative comparison.** Standard metrics are on VGGSound; APCC-Δ is on VGG-Impact. ∗: public-code reproduction. †: author samples. ◇: no test-time text.

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 (Xing et al., 2024)*	415M	0.566	219.0	24.58	5.40	2.26	2.30	8.58	33.99	1.204
V-AURA (Viertola et al., 2025)*◇	695M	0.654	218.5	14.80	2.88	2.42	2.07	10.08	27.64	0.654
VATT (Liu et al., 2024)†	–	0.673	131.9	10.63	2.77	**1.48**	1.41	11.90	25.00	1.195
Frieren (Wang et al., 2025b)†◇	159M	0.662	106.1	11.45	1.34	2.73	2.86	12.25	22.78	0.851
FoleyCrafter (Zhang et al., 2024)*	1.22B	0.588	140.1	16.24	2.51	2.30	2.23	15.68	25.68	1.225
V2A-Mapper (Wang et al., 2024)†◇	229M	0.671	84.57	8.40	**0.84**	2.69	2.56	12.47	22.58	1.225
TARO (Ton et al., 2025)*◇	258M	0.758	159.1	10.49	1.57	2.92	2.67	9.62	22.85	1.169
MMAudio-L (Cheng et al., 2025)†	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

where v_j^- and v_j^+ are pre- and post-impact velocities. This quantity represents mechanical energy lost at impact, which is expected to be radiated as an acoustic impulse (Kinsler et al., 2000). We compare it with the audio onset strength a_j measured from the generated spectrogram (Böck & Widmer, 2013). For each VGG-Impact class, APCC is the correlation between $\{\Delta E_j\}$ and $\{a_j\}$, and APCC-Δ is the absolute gap between the generated and ground-truth correlations averaged across classes. Lower APCC-Δ indicates that generated audio better matches the real coupling between kinetic-energy changes and spectral onset strength.

3.2. Results

Physics consistency. Table 1 shows that existing V2A models often exhibit weak physical grounding even when they produce plausible sounds. Their APCC-Δ values range from 0.536 to 0.758 among strong baselines, indicating a mismatch between visual kinetic-energy changes and generated onset strengths. PAVAS obtains the lowest APCC-Δ of 0.378, reducing the gap by 29.5% relative to MMAudio-L (Cheng et al., 2025), the strongest baseline under this metric. This suggests that explicit mass and velocity conditioning helps the model more closely match the physics-audio relationship present in VGG-Impact. Overall, while existing V2A models may capture semantic alignment, their generated audio only partially captures variations in underlying physical magnitudes.

General audio quality. PAVAS also improves standard V2A metrics. On VGGSound, PAVAS-L achieves the best FD_{PaSST} and FD_{PANNs} among evaluated methods while maintaining competitive synchronization. Compared with MMAudio-L, PAVAS reduces FD_{PaSST} from 60.60 to 47.38 and FD_{PANNs} from 4.72 to 3.99, while increasing ImageBind similarity from 33.22 to 35.41. These gains indicate that physics conditioning does not trade off against perceptual quality; instead, better modeling of object dynamics can improve the generated audio distribution. Table 1 reports the large 44.1kHz variants of MMAudio and PAVAS, while other baselines use their released or author-provided settings. Parameter counts exclude pretrained encoders, latent audio encoders or decoders, and modules not used at test time, following the standard evaluation protocol.

Subjective evaluation. We further conduct a user study on VGGSound clips in which 27 participants rate eight generated audios on audio quality, semantic alignment, temporal alignment, and physical plausibility using a five-point Likert scale (Likert, 1932). PAVAS-L receives higher mean ratings than MMAudio-L on all four aspects, including physical plausibility (4.37 vs. 3.90), temporal alignment (4.45 vs. 4.06), and semantic alignment (4.47 vs. 4.14). This supports the quantitative trend: explicit physics conditioning is perceptually noticeable, not only reflected in proxy metrics.

Ablations. We analyze whether improvements come from physical conditioning or simply from additional training using the S-16kHz backbone under matched training settings. Fine-tuning the backbone longer without physics cues does not improve VGGSound metrics (FD_{PaSST} changes from 70.19 to 71.99), suggesting that the gains are not explained by extra optimization alone. Conditioning on mass or velocity individually improves FD_{PaSST} to 66.89 and 67.22, respectively, while combining both yields 65.67. Finally, residual Δ-modulation outperforms direct summation (65.67 vs. 67.31 FD_{PaSST}), supporting the choice to inject physics as a controlled residual modulation rather than a direct feature merge.

Scope and limitations. PAVAS involves several pretrained vision modules and is not a simple replacement for future jointly optimized physical perception. The present model focuses on mass and velocity; future work may explore richer physical factors such as explicit material modeling.

4. Conclusion

We presented PAVAS, a physics-aware V2A framework that estimates object-level mass and velocity from video and injects these cues into a latent diffusion generator. By pairing this model with VGG-Impact and APCC, we show that explicit physical conditioning improves perceptual quality and physics consistency over existing V2A baselines. Future work can explore jointly optimized physics estimators and richer physical factors such as explicit material modeling.

5. Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. RS-2025-25441313, Professional AI Talent Development Program for Multimodal AI Agents; No. RS-2024-00457882, National AI Research Lab Project), the National Research Foundation of Korea(NRF) grant (No. RS-2024-00453301), and the InnoCORE program of the Ministry of Science and ICT(26-InnoCORE-01) funded by the Korea government(MSIT).

References

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Böck, S. and Widmer, G. Maximum filter vibrato suppression for onset detection. In *the 16th International Conference on Digital Audio Effects (DAFx-13)*, 2013.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020.

Cheng, H. K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., and Mitsufuji, Y. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. In *ICASSP*, 2020.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Fujisaki, W., Goda, N., Motoyoshi, I., Komatsu, H., and Nishida, S. Audiovisual integration in the human perception of materials. *Journal of vision*, 2014.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Iashin, V. and Rahtu, E. Taming visually guided sound generation. In *BMVC*, 2021.

Iashin, V., Xie, W., Rahtu, E., and Zisserman, A. Synchformer: Efficient synchronization from sparse cues. In *ICASSP*, 2024.

Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., et al. Openclip. *Zenodo*, 2021.

Jeong, Y., Kim, Y., Chun, S., and Lee, J. Read, watch and scream! sound generation from text and video. In *AAAI*, 2025.

Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps: Generating captions for audios in the wild. In *NAACL*, 2019.

Kinsler, L. E., Frey, A. R., Coppens, A. B., and Sanders, J. V. *Fundamentals of acoustics*. John wiley & sons, 2000.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM TASLP*, 2020.

Koutini, K., Schlüter, J., Eghbal-Zadeh, H., and Widmer, G. Efficient training of audio transformers with patchout. In *INTERSPEECH*, 2022.

Labs, B. F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Lee, S.-g., Ping, W., Ginsburg, B., Catanzaro, B., and Yoon, S. Bigvgan: A universal neural vocoder with large-scale training. In *ICLR*, 2023.

Likert, R. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Liu, X., Su, K., and Shlizerman, E. Tell what you hear from what you see-video to audio generation through text. In *NeurIPS*, 2024.

Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M. D., Zou, Y., and Wang, W. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM TASLP*, 2024.

Opoku-Baah, C., Schoenhaut, A. M., Vassall, S. G., Tovar, D. A., Ramachandran, R., and Wallace, M. T. Visual influences on auditory behavioral, neural, and perceptual processes: A review. *Journal of the Association for Research in Otolaryngology*, 2021.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Schröger, E., Marzecová, A., and SanMiguel, I. Attention and prediction in human audition: a lesson from cognitive psychophysiology. *European Journal of Neuroscience*, 2015.

Standley, T., Sener, O., Chen, D., and Savarese, S. image2mass: Estimating the mass of an object from its image. In *Conference on Robot Learning*, 2017.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020.

Ton, T., Hong, J. W., and Yoo, C. D. Taro: Timestep-adaptive representation alignment with onset-aware conditioning for synchronized video-to-audio synthesis. In *ICCV*, 2025.

Viertola, I., Iashin, V., and Rahtu, E. Temporally aligned audio for video with autoregression. In *ICASSP*, 2025.

Wang, H., Ma, J., Pascual, S., Cartwright, R., and Cai, W. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *AAAI*, 2024.

Wang, Q., Zhang, Y., Holynski, A., Efros, A. A., and Kanazawa, A. Continuous 3d perception model with persistent state. In *CVPR*, 2025a.

Wang, Y., Guo, W., Huang, R., Huang, J., Wang, Z., You, F., Li, R., and Zhao, Z. Frieren: Efficient video-to-audio generation network with rectified flow matching. In *NeurIPS*, 2025b.

Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., and Yuan, L. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024.

Xing, Y., He, Y., Tian, Z., Wang, X., and Chen, Q. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024.

Zhai, A. J., Shen, Y., Chen, E. Y., Wang, G. X., Wang, X., Wang, S., Guan, K., and Wang, S. Physical property understanding from language-embedded feature fields. In *CVPR*, 2024.

Zhang, Y., Gu, Y., Zeng, Y., Xing, Z., Wang, Y., Wu, Z., and Chen, K. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv:2407.01494*, 2024.