VGGSound*er*: Audio-Visual Evaluations for Foundation Models

Daniil Zverev[*,1] Thaddäus Wiedemer[*,2,3] Ameya Prabhu[2]
Matthias Bethge[2] Wieland Brendel[2,3] A. Sophia Koepke[1,2]

[1]Technical University of Munich, MCML [2]University of Tübingen, Tübingen AI Center
[3]MPI for Intelligent Systems, ELLIS Institute Tübingen

{daniil.zverev,a-sophia.koepke}@tum.de, {thaddaeus.wiedemer,ameya.prabhu}@uni-tuebingen.de

Abstract

The emergence of audio-visual foundation models underscores the importance of reliably assessing their multi-modal understanding. The classification dataset VGGSound is commonly used as a benchmark for evaluating audio-visual understanding. However, our analysis identifies several critical issues in VGGSound, including incomplete labelling, partially overlapping classes, and misaligned modalities. These flaws lead to distorted evaluations of auditory and visual capabilities. To address these limitations, we introduce VGGSounder, a comprehensively re-annotated, multi-label test set extending VGGSound that is specifically designed to evaluate audio-visual foundation models. VGGSounder features detailed modality annotations, enabling precise analyses of modality-specific performance and revealing previously unnoticed model limitations. VGGSounder offers a robust benchmark supporting the future development of audio-visual foundation models.

1. Introduction

Multi-modal foundation models integrating visual and auditory data foster a holistic understanding of audio-visual content. Rigorous evaluation benchmarks have been instrumental in assessing the effectiveness of multi-modal foundation models [7, 9, 10, 14]. To support this, we introduce an enhanced version of the VGGSound dataset [2], a standard audio-visual classification benchmark.

VGGSound suffers from significant issues: We find that its data is inherently multi-label (e.g., a sample might simultaneously be labelled as `playing drum kit` and `playing acoustic guitar` when multiple instruments are present). This challenge is further compounded by partially overlapping classes (e.g., the label `orchestra` often appears alongside individual instrument classes). Moreover, evaluating the contribution of different modalities to the performance of audio-visual foundation models requires modality annotations for each class, as some annotated classes are either not visually present or not audible. We do not remove these samples as inaudible or invisible cues are common in natural videos and critical for multi-modal benchmarking (after all, a self-driving car should not only stop only when it both sees and hears a pedestrian crossing).

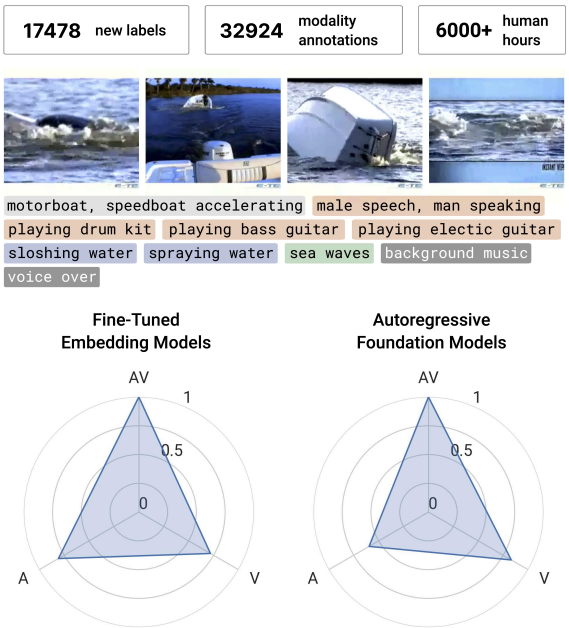

Figure 1. **We introduce VGGSound*er*, a multi-label audio-visual classification dataset with modality annotations**. We extend the `original` VGGSound labels with human-annotated co-occurring `audible`, `visible`, and `visible+audible` classes. We also add `meta` labels for common confounders. Our analysis enables the detailed analysis of auditory and visual capabilities of audio-visual model.

To address these shortcomings, we develop an improved benchmark called VGGSound*er*, following similar works in other domains [1, 5]. We adopt a multi-label classifica-

tion setting by collecting rich annotations on a per-sample basis including (1) additional classes present, (2) explicit modality annotations with each label to mitigate modality misalignment, (3) additional metadata about the presence of background music, voice-over, or static images, and (4) resolve overlapping classes. Overall, we provide a foundation-model-ready benchmark and a structured analysis of whether models rely on audio or visual cues.

We make the following contributions:
1. We illustrate limitations of VGGSound in Sec. 2.
2. We curate VGGSound*er* with multi-modal human annotations for multi-label classification in Sec. 3.
3. We evaluate state-of-the-art audio-visual models, observing differences between embedding models and autoregressive foundation models in Sec. 4.
4. We propose new metrics to quantify modality confusion in Sec. 4.

2. Limitations of VGGSound

Our analysis focuses on the 15.446 10s-long video clips in the VGGSound test set, labelled with one of 309 classes.

Co-occurring classes We find that most samples contain multiple classes (Fig. 2A). These might be temporally separated (e.g., `male speech, man speaking` before cutting to footage of `firing cannon`), or co-occur simultaneously. Overlapping classes are often related, such as different instruments in a band or orchestra, but can also be entirely unrelated.

Overlapping classes Class co-occurrence is exacerbated by many of the 309 automatically generated classes partially overlapping by definition (Fig. 2B). We found two pairs of synonyms: `timpani` vs. `tympani` and `dog barking` vs. `dog bow-wow`. Additionally, some classes are strict subclasses of others, such as the gender-specific versions of `cattle mooing`: `cow lowing` and `bull bellowing`. Finally, many classes commonly appear together, e.g., different musical instruments or semantically similar concepts like `sloshing water` and `splashing water`.

Modality misalignment We find that some classes are invisible or inaudible (Fig. 2C). This modality misalignment is even more pronounced for the numerous co-occurring, unannotated classes: A large fraction of videos contains background music, voice-over or narration, or other background sounds like `bird chirping, tweeting` without a visible source (Fig. 3D). Similarly, some videos contain visible but inaudible cues for classes like `sea waves`. Static images are other frequent sources of misaligned modalities. Finally, some classes are misaligned by definition: `wind noise` is always audible and invisible. We estimate that 48.43% of original samples have misaligned modalities. Visually Aligned Sounds [3], Visual Sound [18], and VGGSound-Sparse [8] removed samples with misaligned modalities. In contrast, we posit that inaudible or invisible cues are common in natural videos and should be considered during benchmarking.

Takeaway 1 VGGSound suffers from class co-occurrence, overlapping class definitions, and modality misalignment, see Fig. 2

3. Building VGGSound*er*

We propose a series of fixes for VGGSound's issues, resulting in the updated VGGSound*er* benchmark.

First, we switch to a multi-label classification setting. This effectively handles co-occuring classes and most overlapping class definitions: A well-performing model can assign a high probability to multiple classes, even if they partially overlap. This also allows us to ensure that synonymous classes and subclass-superclass pairs always appear together in the ground-truth labels.

To deal with modality misalignment, we annotate each label's modality and add meta labels for `background music`, `voice over`, and `static image(s)` to optionally treat these cases separately during evaluation.

Collecting proposals We create a *gold standard* reference set by letting four computer vision experts label a random subset of 417 VGGSound-Test samples that contains each class. We merge these labels via majority vote. Next, we combine predictions from several state-of-the-art models with a manual heuristic to obtain 93% recall relative to the gold standard. We run this classifier on the whole test set to obtain an average of 30 label proposals per sample.

Human labelling We use Amazon Mechanical Turk to re-annotate the entire VGGSound test set and validate the original VGGSound labels with two pipelines; In the first run, annotators enrich the original labels with modality annotations, and in the second run indicate whether the video contains `background music`, `voice over`, or `static image(s)`, then decide for each label proposal whether the class is `audible` and/or `visible` and add missing classes. For both pipelines, we label the samples in batches of 20, each containing two gold standard samples as catch trials. We reject and re-annotate all batches with a catch trial F_1-score below 25% and merge labels via Dawid-Skene algorithm [15]

Final labels Our final labels merge the modality enhanced original VGGSound labels with our annotators' new labels. We automatically add synonymous classes and superclasses for given subclasses. E.g., we add `cattle mooing` if `cow lowing` is in the set of labels.

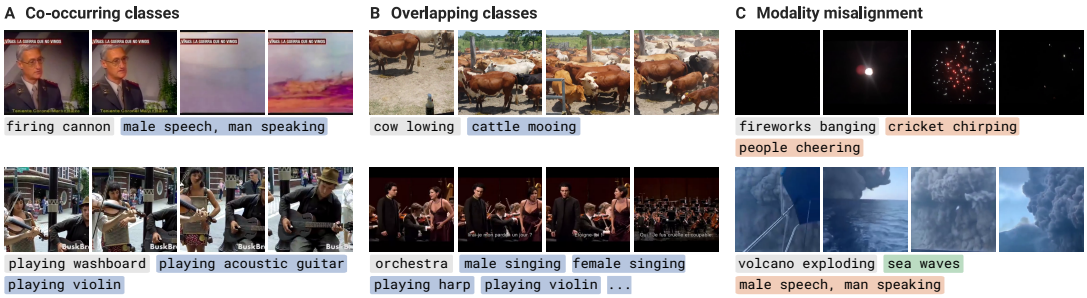

Figure 2. **Limitations of VGGSound.** We show frames of test samples with `original` and `missing labels`. **A**. Many videos contain multiple distinct classes. **B**. Classes often overlap or are ambiguous. **C**. Classes might be `only audible` or `only visible`.

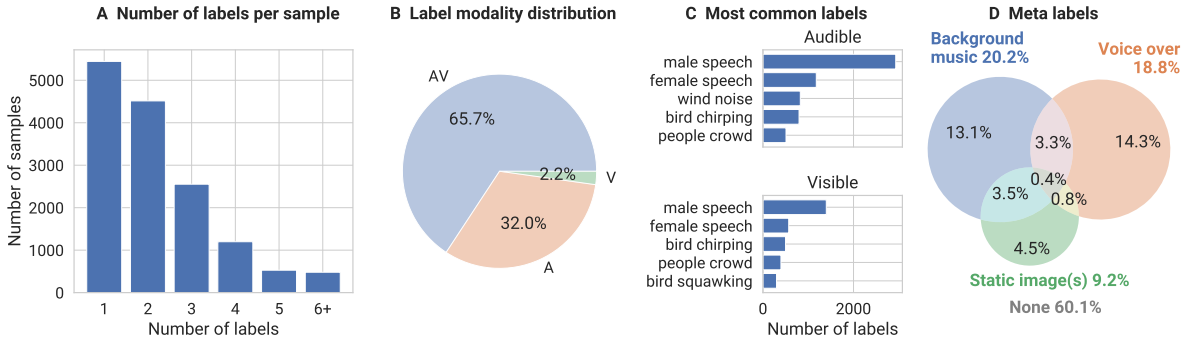

Figure 3. **Overview of VGGSound*er***. **A**. Most samples contain more than one label. **B**. More than a quarter of labels are audible but not visible. **C**. Most videos contain speech. **D**. 40% of the samples contain `background music`, `voice over`, and `static image(s)`.

Takeaway 2 VGGSound*er* extends VGGSound with human-annotated multi-labels, modality annotations, and meta-labels, as summarised in Fig. 3

4. Benchmarking audio-visual models

We benchmark ten audio-visual embedding and foundation models on VGGSound*er*.

Models All models are evaluated in three configurations: unimodal-audio, unimodal-video, or multi-modal. Following [7], we use embedding models pretrained on AudioSet [6] and fine-tune them on VGGSound using the corresponding modalities. Closed-source foundation models are evaluated in a zero-shot classification setup, open-source foundation models undergo an LLM-assisted evaluation protocol [4, 13].

Metrics We use several multi-label classification metrics.

Subset accuracy compares the predicted label set to the ground-truth label set and reports the fraction of samples for which they match. This is our strictest metric.

F_1*-score* is the harmonic mean of precision and recall. It is strictly larger than the subset accuracy.

Hit reports the fraction of samples for which *any* of the predicted labels are part of the ground-truth label set. This is the most lenient metric and strictly larger than the F_1-score.

Modality confusion (μ) is a new metric, which we define as

$$\mu_{\text{modality}} = 100 \cdot \frac{N_{\text{modality-correct,multimodal-wrong}}}{N_{\text{total}}}, \quad (1)$$

where correct/wrong is determined as in the *Hit* score. μ is the fraction of samples that a model classified correctly using a single modality, but misclassified when both modalities were used together.

Foundation models yield an unordered set of class predictions of varying size, and we compute a single metric using the entire set. Embedding model metrics can be computed for the top-k predictions; we chose $k = 1$ to match the foundation models' median number of predictions.

Takeaway 3 We propose *modality confusion* μ to measure how frequently a model is distracted by an additional input modality; see Eq. (1).

4.1. Re-evaluating the state of the art

We present the benchmark performance of state-of-the-art audio-visual models in Tab. 1.

Overall performance Unsurprisingly, all models perform best with access to both input modalities (*AV*). Overall, the foundation models have reached the performance of the specialized embedding models. However, the embedding models fine-tuned on VGGSound generally have stronger unimodal performance with audio inputs compared to visual

Model	Subset Accuracy ↑			F_1 ↑					Hit ↑			μ ↓		
	A	*V*	*AV*	*A*	*V*	*AV*	*A*¬*V*	*V*¬*A*	*A*	*V*	*AV*	*A*	*V*	*A*∩*V*
Embedding Models														
CAV-MAE [7]	**22.57**	26.58	**33.22**	**40.25**	39.54	**48.87**	**21.04**	31.31	**64.54**	54.81	**67.26**	3.60	5.86	0.78
DeepAVFusion [14]	16.13	15.16	28.56	28.85	23.17	42.33	15.22	16.06	46.36	32.15	58.32	**3.29**	**3.51**	**0.11**
Equi-AV [9]	18.72	14.70	26.87	33.66	23.01	39.79	17.97	17.56	53.98	31.90	54.77	6.71	6.92	1.43
AV-Siam [10]	21.66	**27.34**	30.18	38.85	**40.35**	44.37	20.19	**31.89**	62.30	**55.93**	61.08	9.88	9.07	3.94
Closed-source Foundation Models														
Gemini 1.5 Flash [17]	1.87	17.87	19.00	11.94	38.90	43.38	14.09	28.92	25.23	47.63	58.81	8.33	**4.36**	0.61
Gemini 1.5 Pro [17]	**2.95**	**24.96**	**24.11**	**16.97**	**50.65**	**52.93**	**17.70**	**31.94**	**28.74**	**68.08**	**73.47**	**2.00**	4.96	**0.53**
Gemini 2.0 Flash [17]	2.74	15.62	14.38	11.99	36.30	37.75	8.47	27.49	17.23	44.10	47.94	2.06	5.22	0.93
Open-source Foundation Models														
VideoLLaMA-2 [4]	15.68	**20.51**	23.97	35.50	**43.00**	46.75	21.68	**33.66**	44.14	**39.00**	44.48	9.82	**4.15**	2.21
UnifiedIO 2 [12]	**19.00**	18.58	**34.88**	41.07	36.70	**56.30**	28.67	33.14	**54.97**	35.86	**68.49**	**7.65**	5.71	1.80
PandaGPT [16]	7.43	11.19	12.59	28.20	31.25	33.61	23.48	25.61	27.72	26.36	28.54	10.13	8.81	3.17
Ola [11]	15.35	9.97	20.19	**41.31**	23.99	43.02	**36.94**	19.83	42.06	20.35	40.16	11.08	5.23	**1.65**

Table 1. **Audio-visual video classification results on VGGSound*er***. We report multi-label classification metrics (subset accuracy, F_1-score, hit accuracy, modality confusion μ) for audio (*A*), visual (*V*), audio-visual (*AV*), audio-only ($A\neg V$) and video-only ($V\neg A$) inputs. The embedding models were fine-tuned on the VGGSound training set. The closed-source multi-modal foundation models Gemini and open-source models use a zero-shot evaluation protocol and LLM-assisted protocol respectively.

inputs, which is in line with their pretraining. Interestingly, this trend is reversed for foundation models.

Takeaway 4 *Foundation models* perform comparably to finetuned *embedding models*. Embedding models rely more heavily on audio cues than on *visual* ones, the reverse is the case for foundation models, see Tab. 1.

Modality confusion Looking at the *modality confusion* μ, all models have a substantial fraction of test samples (4-11%) where adding a modality actively harmed performance. Furthermore, for all models, a small portion of test samples is not solvable multi-modally, even though they were solvable in either modality alone ($\mu_{A\cap V}$). This insight is enabled by VGGSound*er*'s per-label modality annotations and shows that all models are susceptible to being distracted by an additional modality. This is a concerning issue for multi-modal models which should preserve unimodal capabilities when adding a second modality. Evaluating this behaviour on the VGGSound*er* benchmark serves as a first step towards enabling the development of mitigation strategies for improved audio-visual models.

Takeaway 5 Our *modality confusion score* μ reveals that *all models* can be negatively impacted by additional modalities, see Eq. (1) and Tab. 1.

Performance across modalities Fig. 4 shows performance profiles across modalities. VideoLLaMA-2's performance is well-balanced, while Gemini 1.5 Flash/Pro distinctly underperforms on audio inputs. Embedding models balance

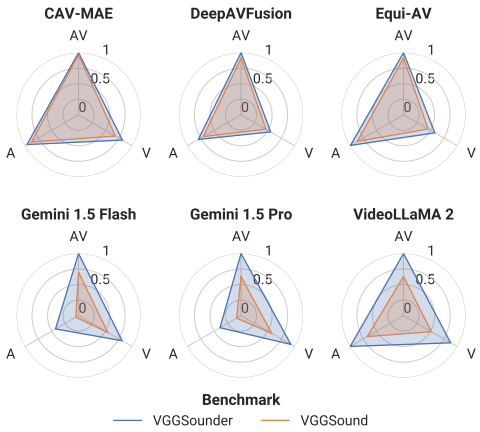

Figure 4. **VGGSound*er* more accurately shows model performance across modalities**, here Hit score on VGGSound*er* and accuracy on VGGSound, normalised by the per-model maximum.

modalities better, with DeepAVFusion and EquiAV slightly underperforming on video inputs.

Takeaway 6 VGGSound*er*'s more complete ground-truth labels allow for more accurate, modality-specific profiling of model performance; see Fig. 4.

5. Conclusion

We present VGGSound*er*, an annotation-rich test set for audio-visual foundation models featuring (1) comprehensive human annotations for missing classes, (2) modality annotations, and (c) meta-labels for frequently occurring real-world challenges.

Acknowledgements

The authors would like to thank Felix Förster, Sayak Mallick, and Prasanna Mayilvahananan for their help with data annotation, as well as Thomas Klein and Shyamgopal Karthik for their help in setting up MTurk. They also thank numerous MTurk workers for labelling. This work was in part supported by the BMBF (FKZ: 01IS24060, 01I524085B), the DFG (SFB 1233, TP A1, project number: 276693517), and the Open Philanthropy Foundation funded by the Good Ventures Foundation. The authors thank the IMPRS-IS for supporting TW.

References

[1] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 1

[2] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 1

[3] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 2020. 2

[4] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 3, 4

[5] A. P. Gema, J. O. J. Leang, G. Hong, A. Devoto, A. C. M. Mancino, R. Saxena, X. He, Y. Zhao, X. Du, M. R. G. Madani, C. Barale, R. McHardy, J. Harris, J. Kaddour, E. van Krieken, and P. Minervini. Are we done with mmlu? *arXiv preprint arXiv:2406.04127*, 2024. 1

[6] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 3

[7] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. Glass. Contrastive audio-visual masked autoencoder. In *ICLR*, 2023. 1, 3, 4

[8] V. Iashin, W. Xie, E. Rahtu, and A. Zisserman. Sparse in space and time: Audio-visual synchronisation with trainable selectors. In *BMVC*, 2022. 2

[9] J. Kim, H. Lee, K. Rho, J. Kim, and J. S. Chung. Equiav: Leveraging equivariance for audio-visual contrastive learning. In *ICML*, 2024. 1, 4

[10] Y.-B. Lin and G. Bertasius. Siamese vision transformers are scalable audio-visual learners. In *ECCV*, 2024. 1, 4

[11] Z. Liu, Y. Dong, J. Wang, Z. Liu, W. Hu, J. Lu, and Y. Rao. Ola: Pushing the Frontiers of Omni-Modal Language Model with Progressive Modality Alignment, Feb. 2025. 4

[12] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action, Dec. 2023. 4

[13] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models, June 2024. 3

[14] S. Mo and P. Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *CVPR*, 2024. 1, 4

[15] V. B. Sinha, S. Rao, and V. N. Balasubramanian. Fast dawid-skene: A fast vote aggregation scheme for sentiment classification, 2018. 2

[16] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai. PandaGPT: One Model To Instruction-Follow Them All, May 2023. 4

[17] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 4

[18] I. Viertola, V. Iashin, and E. Rahtu. Temporally aligned audio for video with autoregression. In *ICASSP*, 2025. 2