Response to the reviewers

Dear Associate Editor,

Thank you for offering us the opportunity to revise our paper "UAVM: A Unified Model for Audio-Visual Learning". Following the reviewers' suggestions, we have made the following revision and explanations to address the concerns of the reviewers.

1. For concerns about the title, we have changed the title from "UAVM: A Unified Model for Audio-Visual Learning" to "UAVM: Towards Unifying Audio and Visual Models" to more clearly state that what we propose is unifying the audio and visual branches of a multi-modality model, not a unified model for all audio-visual learning tasks.

2. For concerns about the claims, we conduct a major revision on the Abstract, Introduction, Experiment, and Conclusion to make our claim better align with the experiments.

3. For concerns about comparative studies, we explain why we mainly focus on comparing with MBT [1]. This is mainly because MBT is a multi-modality model and has SOTA performance on both AudioSet and VGGSound while other works are single-modality models and perform worse than MBT and UAVM. We also better clarified our contribution and novelty in the revised paper.

4. We add a link to our code, it will take some time for the link to work as our institute will review the code release. We update Figure 4 with a more rigorous experiment setting (stratified sampled set for the retrieval experiment), the conclusion has not changed.

5. In addition to the formal black-and-white text version, we also upload a version for review that has changed part highlighted in blue.

We hope the revised manuscript will better suit the IEEE Signal Processing Letters. In the following, we address the concerns of the reviewers point by point (original reviewer comments in black, responses in blue).

Reviewer 1

Reviewer Comment 1.1 — The paper presents an audio-visual neural architecture for sound event detection. The proposed scheme is very insteresting and the way audio and visual streams are processed is very novel. The paper is clear and well written. Resutls are reported on 2 datasets and the paper reports also an interesting ablation study to give insights about the behaviour of the model.

Reply: We thank the reviewer for the positive feedback.

Reviewer Comment 1.2 — My first comment is about the title. I see the Authors' point, but if one claims to be introducing a unified learning framework, I think that there should be experiment on mulitple tasks. In my opinion, the experimental analysis reported in the paper is rich and solid.

But it cannot be generalized an a unified learning approch. So, my suggestion, is to revise the title and the claims in the introduction of the paper, being more task specific.

Reply: We thank the reviewer for pointing this out and we understand the concerns. While by "unified" we meant unifying the audio and visual branches of a multi-modality model, not a unified model for all tasks, we realize that the original title of "UAVM: A Unified Model for Audio-Visual Learning" could be misleading.

In the revised paper, we change the title to "UAVM: Towards Unifying Audio and Visual Models" to more clearly state what we meant by the word "unified". By removing "for audio-visual learning" from the title, we avoid the possible misunderstanding that the proposed model works for all tasks of audio-visual learning. In addition, we also revise the Abstract to better explain that we are unifying the audio and visual branches of a multi-modality model, and we mainly evaluate the model on the audio-visual event classification task.

Reviewer Comment 1.3 — The experimental results show that the proposed UAVM approach performs rather similarly to the unimodal pipeline. There seems to be a marginal improvement but in most cases it is within the variance of the results. In addition, for AudioSet it performs worse than the feature concatenation. On the other hand, the proposed architecture improves noticeably wrt the state-of-the-art. The fact that with a very simple fusion strategy the proposed modal independent model outperfoms the state of the art is a remarkable result in my opinion. But the proposed modality combination is not bringing any improvement. So, as before, I think that the results are very interesting but the message that the paper seems to convey is, in my opinion, misleading. One advantage of the proposed model is that the unified network is smaller. But I am not sure it is something relevant. Authors briefly mention it in the result discussion. If it is relevant, maybe Authors should quantify a bit more.

Reply: We thank the reviewer for pointing this out.

In the paper, we do not claim that UAVM is better than modality-independent models. Instead, we say "Performance-wise, UAVM is similar to a modal-independent model, and outperforms cross-modal attention models on VGGSound." multiple times in the paper. In addition, we say "Transformer models with pretrained frozen features are a strong baseline with low computational cost. As a consequence, UAVM achieves a new SOTA performance on VGGSound" to attribute the strong performance of UAVM to the training pipeline. Actually, the fact that UAVM performs comparably to modal-independent models even when its capacity is very small ($S_{dim} = 16$) is already a very interesting finding, considering the two input modalities are very different.

In the revised paper, we revise the Abstract, Introduction, Experiment, and Conclusion to further clarify these. In addition, following the reviewer's suggestion, we compare the model size in Section III.B, "UAVM achieves almost the same fusion performance when both modalities are input, and even slightly better results when a single modality is input **with only about 76% parameters**".

Reviewer 2

Reviewer Comment 2.1 — The paper was easy to read and follow. The model is explained nicely – simply yet giving all the necessary details. The experimentation section gives all the necessary information, then summarizes the findings concisely and informatively. The ablation analysis was also relevant and interesting to read. The paper follows the best standards of designing the model and conducting experiments.

Reply: We thank the reviewer for the positive feedback.

Reviewer Comment 2.2 — While the citations in general are fine, `https://openreview.net/forum?id=fXorxxbDvO` seems related: it explores the robustness to the missing modality.

Reply: We thank the reviewer for pointing out this related paper [2], we have added this citation it in the revised paper.

Reviewer Comment 2.3 — I was confused about the comment on lines 43-44 that MBT uses pretrained Imagenet features. Isn't it the case for UAVM too as mentioned on page 2 line 52 right?

Reply: The reviewer is correct, both MBT and UAVM are pretrained with ImageNet. We mentioned "MBT is pretrained with ImageNet" just because we wanted to say it is a fair comparison, i.e., we are not comparing a pretrained model with a model trained from scratch. But we realize it can be confusing.

In the revised paper, Section III.B, we change the sentence to "Note this is a fair comparison as both UAVM and MBT use ImageNet pretraining" to avoid confusion.

Reviewer Comment 2.4 — Some of the results show the overlapping confidence intervals. You cannot draw conclusions if either of them is performing better in this case.

Reply: The reviewer is correct that there are overlapping confidence intervals in Figure 2. But when there is an overlapping confidence interval, we only claim the two models are comparable instead of one being better than another.

In the revised paper, we check this throughout the paper and make sure there is no misleading result interpretation.

Reviewer 3

Reviewer Comment 3.1 — It claims that "existing works only unify a part of the model components while the proposed system unifies almost everything." But from Fig.1, the proposed model also seems to need a modal-specific feature extractor and modal-specific Transformer, and only the upper layer shares the weights. In other words, it also looks like only unifying a part of the model; therefore, the novelty is unclear.

Reply: We thank the reviewer for the comment.

The reviewer is correct in that our model also needs modal-specific feature extractors and (optional) modal-specific Transformer layers. But by "existing works only unify a part of the model components while the proposed system unifies almost everything", we meant that existing works do not unify all three of 1) model architecture; 2) model weights; and 3) training algorithm as we mentioned in the second paragraph of the Introduction. Our UAVM unifies all three modal components, which is novel. In addition, we also report a set of intriguing proprieties of the unified audio-visual model, which is also novel.

In the revised paper, we revise the Abstract, Introduction, Experiment, and Conclusion to more clearly state our contribution and novelty.

Reviewer Comment 3.2 — If the model-specific Transformer is only optional, then the experimental results shall justify the performance with and without the modal-specific Transformer.

Reply: The reviewer is correct on this point.

In the revised paper, Section III.B, we add an additional result of a UAVM model with 0 modality-specific layer, which has an accuracy of 65.6% on VGGSound, which is very close to that of the UAVM model with 3 modality-specific layers (65.8%). This justifies our claim that the modality-specific layer is optional, but could marginally improve the performance.

Reviewer Comment 3.3 — Furthermore, it still uses separate audio-based and video-based predictions and then a fusion layer at last. So except for the part that both modalities use the same architecture, both modality seems still work independently. So the statement ”the difference between a unified model and a modal-independent model remains unclear” is unfortunately still valid.

Reply: We thank the reviewer for the comment. For clarification, in our work, the two modalities not only use the same architecture, but also share model weights and are trained with a unified algorithm.

In fact, one major contribution of this work is answering the questions of 1) when unified model and modal-independent models are different? (when S_{dim} is small) and 2) what's the difference? (unified model and modal-independent models have similar performance, but the representation spaces are very different). To the best of our knowledge, these have not been explored before. We hope these findings make the difference between a unified model and a modal-independent model more clear.

In the revised paper, we revise the Abstract, Introduction, Experiment, and Conclusion to highlight our novel findings in the unique proprieties of UAVM.

Reviewer Comment 3.4 — There are other existing works that proposed a unified model, such as the one by Ao et al., ”SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing.” Although the modality is different, I think they may discuss the similarity and originality in terms of the Transfomer-based unified model.

Reply: We thank the reviewer for pointing out this related paper [3], we have cited it in the revised paper.

Reviewer Comment 3.5 — Despite several related works on the unified model introduced in the Introduction (such as [14]-[17]), unfortunately, none of those models are included in the experiments. So whether the proposed method could provide advantages in comparison to existing works is unclear.

Reply: We thank the reviewer for pointing this out. Although [14]-[17] (of the original manuscript) also study unifying the audio and visual branches, they are evaluated on different tasks and/or in different evaluation settings with our work. More specifically, [14] and [15] are not evaluated on AudioSet or VGGSound. [16] and [17] are evaluated on AudioSet and VGGSound, respectively, but only in a uni-modal (audio) setting. In addition, both [16] and [17] audio-based results are worse than MBT [1] and our UAVM. To the best of our knowledge, MBT is the SOTA model on AudioSet and VGGSound, and it is a multi-modality model. We therefore focus more on comparing with MBT and other strongest existing models. The technical difference between [14]-[17] and our UAVM model is mentioned in the Introduction, second paragraph.

Reviewer Comment 3.6 — Compared to the independent model, the performance of the unified model doesn't seem to improve significantly.

Reply: The reviewer is correct on this point. However, throughout the paper, we do not claim that UAVM is better than modal-independent models. Instead, we mentioned that "performance-wise, UAVM is similar to a modal-independent model" multiple times in the paper. Actually, the fact that UAVM performs comparably to modal-independent models even when its capacity is very small ($S_{dim} = 16$) is already a very interesting finding, considering the two input modalities are very different. Besides the classification performance, we also show that UAVM is stronger in audio-visual retrieval.

In the revised paper, we further clarify that the performance of UAVM is similar to modal-independent models to avoid misleading.

Reviewer Comment 3.7 — Figure 3D shows that the shared Transformer maps two different inputs to a unified space. But it is still unclear whether both modalities share the same space but are still independent or both modalities also have a similar distribution. For example, will the map show the data in a close location if audio and video have the same content?

Reply: We thank the reviewer for the comment. We did exactly what the reviewer asks in Section III.D "Audio-Visual Representation Correspondence". More specifically, we evaluate the audio-visual retrieval performance to show the distance between paired audio and visual inputs is indeed closer than unpaired audio and visual inputs in the representation space. And UAVM is better to capture such audio-visual correspondence than modality-independent models.

In the revised paper, we revise the first paragraph of Section III.D to explain the motivation of this experiment.

References

[1] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, pp. 14 200–14 213, 2021.

[2] O. Chang, O. Braga, H. Liao, D. Serdyuk, and O. Siohan, "On robustness to missing video for audiovisual speech recognition," *Transactions on Machine Learning Research*, 2022.

[3] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," in *ACL*, 2022.