We thank reviewers for their thoughtful feedback.

Additional references (R1, R2, R3). We reported in Table 2 the best existing self-supervised baselines at the date of submission (including unpublished methods such as XDC). We outperform all the reviewers' additional references that target unsupervised learning (R2's [1,2]), apart from results on a single benchmark (ESC-50) for R2's [1] which we indeed miss and will add (it appeared on arXiv one month before the deadline and is unpublished). Other references are relevant but do not address the problem of unsupervised representation learning (e.g. "One model to learn them all" is about supervised learning, "Multimodal transformer for unaligned multimodal language sequences" uses supervised features...). We will add all mentioned references but we stress that they do not hurt our novelty claim nor our results.

Fair comparison to previous work (R2, R4). Comparison on equal grounds is a problem for all papers in this area, and we try to be fair: (1) The **comparison with MIL-NCE (R4)** is performed in Table 1(a) since MIL-NCE is strictly equivalent to our VT only architecture. With the exact same training setup and backbone, our FAC outperforms MIL-NCE by 2% in both UCF and HMDB, while enabling a task (ESC-50) that would not be possible with MIL-NCE. (2) The point of the paper is to demonstrate that **more modalities help (R2)**. Note that we outperform ELo which uses an additional modality that we don't (flow, which is really important for action recognition). (3) It is difficult to compare our method on the **same data and architectures** as (i) FAC requires text and only HowTo has it (only MIL-NCE uses this dataset, and we compare to this), (ii) IG65M is not public (**R1**), (iii) AVTS, XDC and ELo haven't released code, and (iv) other works use a variety of architectures. We beat XDC despite using less data (16 vs 21 years) and fewer parameters (24 vs 33M), and ELo despite using a $2\times$ smaller model. (4) We did **another comparison to XDC (R2, R4)** by running our VA model on the same data (AudioSet) and backbone (R(2+1)D-18) and outperform them: 91.5 (vs. 91.2) on UCF and 70.1 (vs. 61.0) on HMDB (also matching R2's [1] on UCF and beating on HMDB on the same data). Note that R(2+1)D-18 actually outperforms S3D (used in our submission) so MMV does not simply beat XDC due to a better backbone (**R4**). (5) **XDC beats their own fully supervised baseline** but we report a stronger and more meaningful quantity – the best externally published performance for supervised transfer. Finally, as detailed next, we would like to stress that our good performance (e.g. a significant boost of **5.5** point on HMDB) is not the only contribution of the paper.

Novelty, contributions and claims (R2). We agree that our novelty does not lie in the loss which is indeed not novel. The loss is not the only means to induce different structures on the embeddings, instead, we achieve that through architecture design. For example, the FAC design allows us to navigate from the video-audio space (fine) to the video-text space (coarse) (property (iii) of the MMV L28), which is not possible with the disjoint design. We validate that claim in the supplemental video as highlighted in L313 through qualitative audio-to-text retrieval (we are not aware of standard benchmarks for quantitative audio-text retrieval evaluation). This shows that we can influence the structure of the embedding through architecture design and not only with losses. Furthermore, we show that FAC performs better than the other considered designs in Table 1(b) on 3 out of 5 benchmarks. In addition, to the best of our knowledge (acknowledged by R1 and R4), this is the first work to jointly learn from video, audio and text in a self-supervised fashion. In the paper, we explore how to do that well at scale, propose various embedding strategies, and demonstrate state-of-the-art performance on challenging downstream tasks which we deem to be an impactful contribution.

Importance of the deflation contribution (R3). We believe the deflation technique is an important contribution of the paper as it allows video-trained models to do inference efficiently with image inputs. In particular, we show that the image classification performance when using the deflated model is similar to using the original model with an inflated input. In addition, we believe to be the first work to consider learning first from video to transfer to images and show strong performance on two image benchmarks. We will clarify the impact of the deflation contribution in the paper.

Baseline for deflation (R1). R1's proposed method is indeed a valid idea which we expect to perform on par or better than our approach. However, this method effectively does partial finetuning, while we instead focused on linear evaluation with frozen networks since it enables fast off-the-shelf evaluation on unseen image dataset and tasks.

Positives and negatives for NCE (R1, R3). For the text positive (L180), we strictly follow MIL-NCE [41] and refer to this paper for details. In L173, we use all negatives coming from other elements in the batch. In total there are $2 \times (N-1)$ negative pairs ($N-1$ audio negatives for the video and $N-1$ video negatives for the audio).

Model and code release (R2, R3). We will release our pretrained models as well as the training scripts.

Perfect audio-visual alignment (R3). We agree there is no perfect alignment, but the point of L174 was to emphasise that text is less aligned with the visual content than audio in general. Note that all competing audio-visual learning approaches ignore the occasional misalignment and, like us, observe that learning is possible despite it. We will clarify.

ASR for action recognition (R1). Text from ASR provides semantic information about the visual content of the videos for objects (e.g. tomato) or actions (e.g. cut). Interestingly even though there exists a domain gap between the