We thank the reviewers for their thorough evaluation. We believe that the reviews will help us to clarify and improve the quality of the submitted paper.

Our main contribution is to use contrastive learning for creating image-like embeddings suitable for registration, and modifying the InfoNCE loss to obtain the rotational equivariance property. This framework can be used with any critic, and we empirically show that MSE leads to good results for the registration task in particular. InfoNCE has been previously used to learn embeddings used in classfication and segmentation tasks in which the resulting subspace is required to feature properties such as separability between classes. However, we are the first ones to produce image-like, contrastive representations that possess the necessary equivariant properties to find a transformation (through classical registration methods, tested with rigid models) between the original inputs.

Our method requires aligned pairs of images available for training, but the registration algorithm applied to the CoMIRs can be chosen independently and could use affine or deformable models. We test rigid models in this study and outperform [29], the SOTA of intensity-based, affine registration based on biological properties between the very dissimilar modalities SHG and BF. While registration under diffeomorphic transformations is a very interesting problem to address on top of our current experiments, our comparison with the baseline and SOTA method [29] shows that rigid registration using rotations and translations is a very challenging task on this biomedical multimodal dataset (see Fig. 4), due to the little correlation between the modalities. Apart from comparing with [29], we also mention [22] and [25] as other works of the medical community. We acknowledge MIND, which is one of the pre deep-learning approaches, and use mutual information (non learning-based method) as a baseline. The proposed references seem relevant to our work, and could be added after an in-depth reading.

We show that data augmentation alone does not achieve rotational equivariance, but that equivariance is reached by modifying the loss (Fig. 2, augmentation is applied to both C_4 enabled and disabled experiments). The cost of data augmentation and equivariance enforcing components are independent of the number of dimensions; in 2D, for every step we sample one rotation from a set of four (0, 90, 180, 270 deg.), in 3D, we would still sample only one rotation but from a set of 24 rotations. The pipeline may have to train longer for the networks to experience the CoMIRs rotated in all possible directions sufficiently many times, but considering the observed fast convergence (23 epochs or 22 minutes [l. 566 & App. Table 3] for the biomedical dataset) we believe this is feasible. We show that multiples of 90°provide nearly perfect rotational equivariance (Fig. 2 & `https://youtu.be/SRISQFfOVI4`) where we ablate the component enabling the C_4 equivariance. It is reasonable to expect that the minute drops in performance around multiples of 45°(Fig. 2, polar plot) disappear by using the C_8 group, but this relatively small improvement comes at the cost of interpolating the CoMIRs for both the forward and backward passes.

The reviews contain many suggestions on how to clarify and improve the article. They are all very relevant, and we appreciate and fully agree with those comments.

Reviewer 1: The main computational cost of the method is linear w.r.t. modalities, as only one model is trained per modality. However, some components of the loss grow quadratically (the similarity matrix is $MN \times MN$, with M modalities and batch size of N) but are inexpensive to compute, in comparison to a forward pass of a model. As the other modalities are being used as negatives, doubling the number of modalities and dividing the batch size by 2 results in the same computational cost. We consider the critic function as a hyperparameter, which should be chosen according to the task at hand [53] and can differ from cosine similarity, as usually chosen for unsupervised classification. We hypothesize that the MSE encourages the intensities of the representations to be more similar than cosine similarity, which is favorable for registration. In the future, we would like to further explore and understand the effect of the choice of critic. We thank the reviewer for advising us to explore gCCA, which seems highly relevant.

Reviewer 2: We are thankful for the detailed comments regarding notations and will clarify the equations and text accordingly. The details regarding the MI bound for $n \to \infty$ are given in [45], App. A1. The question regarding what pixel error is acceptable is highly relevant. As the ground truth was obtained by manual registration, we included an independent manual registration task of a subset of our experimental setup by 6 human annotators which showed that a pixel error up to ~ 50 pixels can be expected (see App. Fig. 7).

Reviewer 3: Both datasets are publicly available for the community ([1] & [56]), to reproduce and compare our method and results. The best performing registration method using CoMIRs has a 3 times higher success rate, resulting in less than 1% pixel error, than the SOTA on the given modalities (SHG and BF, not using CoMIRs) and more than double the success rate for a given tolerance of 10% pixel error (Fig. 4). We experimented with different contrastive approaches (triplet loss and multi-class n-pair loss [48]) and InfoNCE empirically performed the best in terms of stability and suitability of the representations for registration. Therefore we focused on InfoNCE for this study.

Reviewer 4: We compared our approach with approaches that do not require paired images (MI and GANs) and showed that they were outperformed by our method. In cases where aligned pairs cannot be obtained by acquisition, manual registration or computational means, non learning-based approaches such as MI remain the best option available.