## Fair Comparison with Supervised Teacher

Throughout the experiments with the end-to-end self-supervised approach, we observed significant improvements in performance. This increase is especially notable when comparing the results with the supervised teacher used at the beginning of the experiments on multimodal knowledge distillation. However, the comparison with our Transformer-based SHRe model "$\text{SHRe}_\text{T}$" is not entirely fair, as the self-supervised teacher BEiTv2 is a much larger model compared to the ResNet-50-A1 teacher of "$\text{SHRe}_\text{T}$": BEiTv2 has 86M parameters, while ResNet-50-A1 has only 25M parameters. To have an actual comparison between a self-supervised (our approach) and a supervised (SHRe approach) teacher, we:

1. Add all improvements that are not specific to the self-supervised approach to "$\text{SHRe}_\text{T}$".
2. Train $\text{SHRe}_\text{T}$ with a supervised teacher that is comparable in size to BEiTv2.

As for the first point, we add the improvements from the self-supervised approach to "$\text{SHRe}_\text{T}$". This includes the following:

1. ...

For the second point, we keep BEiTv2 as the teacher, but now use the variant that has not only been trained self-supervised on ImageNet-1K, but also finetuned on ImageNet-1K with labels, i.e., the supervised variant of BEiTv2. That way, we can observe the direct impact of switching to a supervised teacher.