Response

Sincere thanks to the reviewers comments and suggestions. And thanks to the senior area chair for reading our response.

1. First, we would like to clarify two factual errors made by Reviewer HjPV.

- Reviewer HjPV stated that "The basic hypothesis that ..... major weakness of this paper." and "The entire image pool ..... how can we index the image set?". Both complaints are not true. First, the development of learner models while maximally retaining performance is always useful for resource-constrained scenarios including servers with limited budget, embedded systems, and mobile devices. If not, what is the point of all previous researches on model compression? Second, Reviewer HjPV ignores the important scenarios where users do not want to use the cloud to store their photos due to privacy and economic concerns, but instead prefer a local solution that allows fast and accurate search over their private collections.

- Reviewer HjPV rated 1 for the software availability. This is wrong as we have attached the software to the ARR system. We also included the anonymous URL to the repository containing the code, models, and demo application. The other two reviewers both acknowledged the useful software and rated 4.

2. Second, we would like to respond to four critical comments raised by the reviewers and meta-reviewer.

- Reviewer HjPV stated that "Relevance of this paper for ACL community is also questionable". We disagree. On the one hand, we have seen at least 7 papers on cross-modal retrieval that were published at *ACL conferences just in 2021. On the other hand, the official ACL CFP specifically includes topics such as 1) Information Retrieval and Text Mining and 2) NLP Applications.

- We appreciate that Reviewer EeRM checked the CLIP paper when preparing their review, but we would like to clarify that the results reported in the original CLIP paper correspond to the largest version of CLIP. This version is not open-sourced, and cannot be used as our compression target. The results in our paper correspond to the second largest version of CLIP, which is publicly available.

- Reviewer EeRM suggested fine-tuning ViT-small and TinyBERT as a baseline. We did not include it in the current version because we have conducted a similar ablation in Table 2, i.e., stage-1$_{InfoNCE}$. This ablation is almost identical to what Reviewer EeRM suggested except that the text encoder is not TinyBERT. TinyBERT is not pre-trained on massive image-text pairs as is done for CLIP hence we posit that it will not deliver better or even close performance compared to ours. Nonetheless, this baseline is plausible and straightforward to implement, and we are very willing to include it as our baseline in the final version.

- "No novel knowledge distillation technique" comment made by the meta-reviewer contradicts the comments by other reviewers. For example, reviewer HjPV mentioned "the loss function shows novelty", and reviewer EeRM mentioned "An elegant way to compress large pre-trained dual-encoder".