We thank the reviewers for their thorough and very helpful feedback. We are glad that all reviewers found the dataset to be a valuable contribution—we believe that this work is important for providing better measurements for multimodal AI research in the future, with a clear positive contribution to society as a consequence. We address each reviewer below:

Reviewer 1 Thank you for your insightful review, we will do our best to incorporate your excellent suggestions.

We will include a more detailed analysis of the dataset properties in the camera ready, if accepted, including of the dev set and a breakdown of multimodal vs unimodal hate, benign image/text, other random non-hateful. We did not do this initially because we wanted to avoid compromising our "unseen" dataset.

"An additional evaluation [..] using subsets of the training set of different sizes could shed some light" – Thank you for this excellent suggestion! We quickly did this experiment for the MMBT-Grid model and performance goes up considerably from using 10% of the training data (60.46 ROC-AUC on dev) to 50% (64.00) to 100% (68.57) of the training examples. We will include a plot in the camera ready, as well as provide further analysis.

We agree about real world meme generalization. Many such memes do use stock photos, however, and since we also release the raw SVG files it is easy to create different variations of the same meme, which is an interesting research direction. We will also add a column for easy/middle/late fusion to Table 1 to make that clearer.

The unimodal versions of VilBERT and Visual BERT are essentially the initializations used when pre/inter-training ViLBERT and VisualBERT models: rather than first training on multimodal data (e.g., COCO or Conceptual Captions), these models are finetuned directly on the Hateful Memes task without the intermediate training step.

Reviewer 2 We really appreciate your thoughtful review and look forward to incorporating your comments.

We will include a plot of varying training dataset sizes in the camera ready, if accepted (see above). We will also include further analysis of the label quality as it relates to dataset size (our analysis for R1 above showed that even 10% of the training data is very useful, so you make a good point) – thanks for this suggestion. As you note, annotation was very costly, so this trade-off is definitely worth making explicit and examining further.

We agree that using images from a single source like Getty could make the distribution different from (some) real world memes. However, since the same procedure was used for all memes in the dataset, we think that it isn't a huge problem here, especially since many real memes are built using stock images as well. We also release the SVG files, so we hope that future work will try to analyze this further by replacing the background images and modifying the text properties.

An analysis of different model failure modes will be very interesting indeed—from what we have seen, the top models make similar mistakes, which will be useful to demonstrate in-depth, thanks for the suggestion.

Non-standard text is handled by the text-encoders: the transformer-based models all use Byte-Pair-Encoding, which means they are more robust to typographical errors, acronyms and out-of-vocabulary words, but you are definitely right that this would be a good avenue for trying to improve model performance on this task.

Reviewer 3 Thank you for your review. We were a bit surprised by some of your points, which we hope to address:

Regarding the paper's organization: We respectfully disagree with your assessment—in fact the other reviewers all note that the paper is well written. We agree that this paper's contribution is different from more standard dataset papers (which we think is a good thing), which also means that we have to spend more time discussing the non-standard annotation process (i.e. in describing how we define hate speech or how we obtain benign confounders). We will happily include more dataset analysis, and will endeavor to make it even clearer what the dataset improves over previous work.

With regard to the binary label, we believe that this has several important benefits: i) it makes evaluation straightforward, which is important for machine learning problems, especially if we are trying to encourage the community to tackle an important problem together, for the greater good; and ii) as we describe in the paper, a binary label is actionable in practice: if a meme is hateful, it can be taken down; if a meme is disagreeable but ultimately not hateful, it should stay up – this distinction is ill-defined for an alternative finer-grained labelling. We agree that finer-grained labels can also be very valuable and should be investigated, but that question is unfortunately out of the scope of this work.

We respectfully disagree that the baseline models are too simple: we used state-of-the-art multimodal models, which are well-known as such in the V&L community. Note that MMBT, which uses grid features and is much simpler than ViLBert and VisualBert, compared to gated fusion in their paper and beat it; DeepDualMapper is specific to images and does not incorporate textual information. That said, we would happily include gated fusion as well in the camera ready.

Reviewer 4 We thank you for your support and very useful feedback.

You are absolutely right that the benign confounders introduce a slight skew to the source images. Do note that the text will be different in each case, so if anything this skew makes the dataset even more difficult. You make an interesting point however, and we will examine if this has an impact in the camera ready, if accepted.