We thank all reviewers for their detailed and constructive comments. **R1:** "potentially high impact contribution," "techniques and analysis also interesting," **R2:** "problem [...] is relevant and important," "dataset is original," **R3:** "really interesting and challenging problem," "prose is clear and the paper is quite enjoyable," "experiments are plentiful".

All reviewers. Ours-GT: Apologies for the confusion; we will clarify. OURS-GT uses the **G**round **T**ruth text paired with the images *at test time* (to compute a document embedding), in addition to the image. We thus consider it an upper bound to the task of visual only prediction. It is the same as the first stage of our approach, without the addition of the image classifier layer. We will include results for the upper bound in Table 2 as requested by **R2**.

R1: Contribution of stage 2: If we remove stage 2 and zero out weights for text embedding, acc. is only 0.677.

R1: "Sweet spot" for text data: We will include an experiment that trains with the first k sentences (varying k).

R2: "Strong assumption" of no text at test time: As R3 notes, there are numerous works (cited in our main text) which study the problem of predicting political bias from natural language. In this work, we wanted to explore prediction of *visual* political bias. We show that predicting bias from images alone is more challenging than prediction from text. The OURS-GT model uses ground truth text at test time and clearly outperforms all other methods. Our analysis indicates that certain words and phrases are easy "giveaways" of bias: e.g. "fascist" or "Nazi" imply left bias, while use of words like "communist" or "socialist" imply right bias. Thus, while a model which has access to test article text (in addition to the image) performs better at test time, we don't know if the predictions are primarily visual or textual. From a practical perspective, if one only cares about predicting bias as accurately as possible, they can of course use the OURS-GT model, or develop novel ways to best utilize the combination of image and text. This is not our focus. The purpose of our work *from a scientific perspective* is to study purely *visual* political bias. We further show how text can be leveraged in this context, while still enabling visual-only prediction.

R2: Problem-specific approach of leveraging text as privileged information: We refer R2 to Gomez et al., "Self-supervised learning of visual features through embedding images into text topic spaces", CVPR 2017, which uses an approach trained to predict text embeddings from images. The features are then applied on visual-only data, e.g. PASCAL image classification. We tried a variant of this approach of predicting latent text topics from images and obtained 0.681 on our full dataset (much lower than our method; compare to Tab. 1 in main). Predicting text embeddings from images is too challenging on our data because of the many-to-many relationship of images w/ topics (e.g. image of the White House can be paired with text about Trump's children, border control, LGBT rights, etc.).

R2: Two-headed model / DeVise / Word prediction: We experimented with a two-headed model which predicted both bias and the top-1k visual words (see L310 for addl. details) from the first two sentences (these had best overlap with human chosen aligned text). The model achieved 0.626 acc. for bias prediction overall (much lower than our method; compare to Tab. 1). Further balancing of loss hyperparameters could potentially improve the result. We also trained a model that predicted the top 1k visual words from images and then used the word predictions for bias prediction. This achieved 0.567 acc. Thus predicting words or embeddings is less promising than our approach.

R2: JOO performance on "No people": We agree JOO's performance in Tab. 2 is counterintuitive. Note the result on "No people" is statistically equivalent with our method's performance (McNemar's Test, $p \leq 0.05$). Further, JOO's method does use scene context features in addition to person features (see L268-269). Finally, JOO's dataset focused on the 2012 election, while ours primarily deals with 2016, and lacks many of the politicians that appear in ours.

R3: Per-source train splits: Because media sources republish images from others for commentary, it is difficult to ensure *all* images from a source have been excluded. Still, we experimented leaving out all training data harvested from a few popular sources. The result was (before excluding $\rightarrow$ after excluding): DemocraticUnderground (0.713$\rightarrow$0.700), DailyCaller (0.703$\rightarrow$0.667), NewsMax (0.685$\rightarrow$0.628), TheBlaze (0.746$\rightarrow$0.742), Breitbart (0.607$\rightarrow$0.566), CommonDreams (0.647$\rightarrow$0.636), and CNN (0.873$\rightarrow$0.866). We observed only a slight decrease for all sources we tested.

R2-R3: Model details / fusion: The fusion layer is a linear layer which receives concatenated image and text features and produces activations used by the classifier. Thank you, we will clarify in our main text.

R3: Failure cases / model interpretability: We agree that failure cases could be useful and will include some in supp. We will also include activation heat maps over these images to understand what the model used to make its prediction.

R3: "Useful" bias: We agree that the model may be learning things like photo of assault weapon implies left. However, Figs. 4-6 in supp. show that this is also how *humans* approach this task for ambiguous images (e.g. "guns and flag = R").

R3: "Neutral" images: This is an interesting idea, but we consider it orthogonal to our work. Moreover, even agreeing on a definition of neutral sources is difficult, e.g. people disagree on whether The New York Times is biased.

R3: Novelty wrt privileged info: We are not aware of a method like ours. Other approaches use tied weights [4], computing summary statistics [56, Lambert CVPR 2018], or multitask learning [Elliott IJCNLP 2017].