We thank all reviewers for their feedback. We are happy the reviewers agree that our work is novel, insightful and offers a new perspective on bias in multi-modal problems.

Regularization and early stopping [R1,R3]: On the right we present results for early stopping and ℓ_p regularization on the SocialIQ dataset. The baseline is described in the appendix. Our regularization performs better than classical regularization and early stopping.

	epoch 5	epoch 10	epoch 15	epoch 20
Baseline	**66.39%**	62.33%	63.91%	62.78%
Baseline+ℓ_2	66.02%	65.27%	64.98%	65.42%
Baseline+ℓ_1	65.15%	64.24%	62.44%	64.02%
Baseline+ℓ_∞	63.23%	64.13%	63.01%	64.58%
Baseline+Ours	66.16%	**68.08%**	**67.51%**	**67.29%**

Tab. 1 & baselines [R2,R4]: In retrospect, Tab. 1 is confusing as the baselines are different for each task. The baselines are described in the appendix, Section 4. The VQA-CPv2 baseline is based on [23]. The SocialIQ baseline follows [7]. The Dogs&Cats baseline is ResNet18. Baseline** corresponds to these baselines augmented with weight-decay (ℓ_2 regularization). Lastly, max vs. convg is also confusing: we used it to emphasize the inconsistent behavior of ColoredMNIST. We attribute it to the synthetic nature of ColoredMNIST. We'll clarify.

Prior art [R2,R3,R4]: We'll add a comparison to REPAIR on our setting for ColoredMNIST: Our de-biasing achieves 96% accuracy, while REPAIR achieves 84.33%. We compared our performance on Dogs&Cats to "learning not to learn" [6], see L263-L265: for TB1 we got 94.71% and for TB2 we got 88.11%. [6] obtains 90.29% for TB1 and 87.26% for TB2. We'll update to the best reported accuracy on SocialIQ [7] which is 64.82%, while our method improves accuracy to 67.93%. VQA-Rephrasing: the LMH [23] baseline obtains an accuracy of 49.23%, while our regularization improves accuracy to 51.18%.

VQA-CPv2 result interpretation [R2,R3,R4]: Great suggestion to study the differences of VQA-CPv2 question-type results of different models. We don't think we can conclude that one model is better at leveraging high-level image information than another. E.g., 'does the,' 'is the person,' 'are these,' questions are very similar in spirit to 'does this,' 'is this person,' 'are they,' questions: both triplets require intricate image understanding. We improve results on the former three while accuracy drops on the latter three.

R1: *Duplicated, subset and corrupted signals:* Thanks for these suggestions. The relevant plots show that our regularization reduces the amount of information from corrupted signals, while improving accuracy:

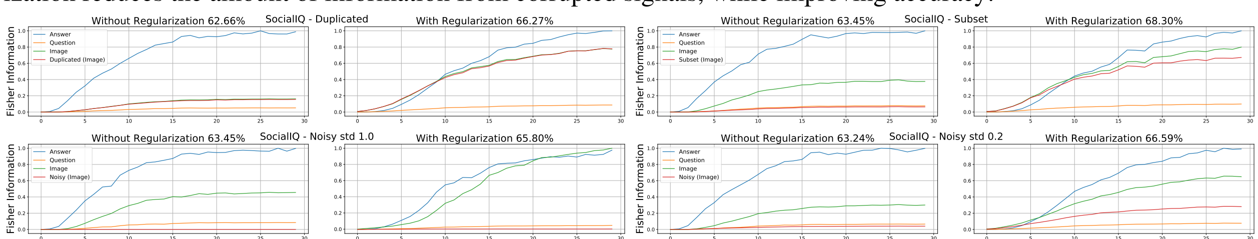

Figure 1: Duplicated, subset, noisy (with different noise levels) modalities: Fisher information (y-axis) as a function of epoch (x-axis) with and without regularization. Accuracy is provided in the plot title. Noisy image is a Gaussian noise added to the image modality.

Bound tightness: The bound is tight for the exponential function $f(z) = e^{tz}$. Since we are using the CE similarity measure over exponential families (through the softmax), our bound tends to be tight. *Modalities overfit at different rates (Wang2020):* Thank you for pointing out this interesting work. Different from our work, this work regularizes the overfitting behavior of different modalities. We'll cite and discuss this work.

R2: *Functional entropy literature:* We acknowledge, finding [32] is not easy due to Covid19, as access to academic libraries is limited. Relevant definitions are also in `https://arxiv.org/pdf/math/0609050.pdf`, Sec. 6. *Clarity:* We'll fix and clarify these 7 points: **1)** The bar plots show the functional Fisher information values; **2)** Answer and question are considered as a "modality" in many VQA works [11, 13, 15, 16, 17, 18]. We wanted to be consistent with prior work; **3)** The relation between Eq. (17) to Eq. (18) is indicated by Eq. (4); **4)** VQA-CPv2 is inherently about debiasing and we compare our method to 5 different debiasing models on VQA-CPv2 in Tab. 2. We also compare to the debiasing work "learning not to learn" on Dogs&Cats in Sec. 5.4; **5)** We detail the settings of each model in the appendix (Sec. 4); **6)** We obtain Fig. 2 by computing the functional Fisher information using Eq. (16) for each data-point and then average over all data-points. In Fig. 3 we use Eq. (2) for 'Ent' and Eq. (3) for 'Var'; **7)** We'll add the citations. *SocialIQ A2 and A4:* We evaluated A4 with a similar model to A2. Our regularization improves the accuracy in this task as well: we obtain 56.35% accuracy without our regularization, and we get 57.13% with our regularization. *Different models for the same task:* Thanks for suggesting, we ran SCR [25] on VQA-CPv2 with our regularization and obtained an accuracy of 49.4%. Without our regularization, we obtain 48.8%. We'll add more models on VQA-CPv2 for the camera-ready. *Answer modality bias in SocialIQ (L243):* We noticed it while experimenting. We clarify and provide the code. *Upweighting regularization term:* When upweighting λ the modalities tend to increase their functional Fisher information at the expense of accuracy. We'll add plots and a discussion.

R4: *Results on VQA v2:* Thanks for pointing out. We used LMH [23] with our regularization and obtain an overall accuracy of 57%, 'yes/no': 66.62%, 'number': 37.97% and 'other': 54.74%. LMH accuracy without our regularization is 56.345%, 'yes/no': 65.057%, 'number': 37.631% and 'other': 54.687%. We obtain consistent improvements. *Focus on softmax function:* as mentioned in L108 the only constraint on f is non-negativity. It can hence be applied to BCE. Note, BCE can be reduced to CE via a binary softmax probability.