We thank the reviewers for the thoughtful feedback! We are encouraged that all voted to accept, finding the DwD task interesting [R1,R2], novel, and compelling [R1,R4]; our approach elegant [R4] and interesting [R2]; and our experiments comprehensive [R1]. [R3] appreciates the use of human evaluations to decisively measure language interpretability. We respond to select comments below but will address all feedback.

[R1] Why not use a more recent VQA model? Performance would improve. Yes, it likely would; however, this is orthogonal to our primary investigation. The focus of our work is on adapting Q-bot's questioning strategy to a dialog without having seen dialog during training. The BUTD model is a well-established model to demonstrate this on. We agree doing so with more recent transformer-based models is an interesting future direction.

[R1] Why no human experiments evaluating human & Q-bot pairs on game performance? These experiments were reported in Sec 4.4. We paired humans with Q-bots and ran the game with humans responding to Q-bot's questions as the reviewer describes. As in L284, game performance with a human answering Q-bot's questions is 69% for our method, 45% for typical transfer, and 23% for zero-shot transfer. This result is a key finding of this work and highlighted in Contribution 3 in the introduction. We also had the human players evaluate the fluency and relevance of questions.

[R2] Game design includes access to the image pool for the questioner unlike in visual dialog. We see this as a strength not a weakness. A question may be discriminative in one pool, but not in another, so questions should depend on the pool. By adopting a pool-conditioned setting, we can evaluate Q-bot's adaptation to different pool sizes, image domains, and pool selection strategies. In contrast, pool-free methods in prior work will always produce the same question for an input regardless of the pool once trained – implicitly conditioning on the training pool. We also note that past pool-free work has found that access to the caption results in the subsequent dialog not playing a significant role.

[R2] The model is trained on VQA data that only focused on discriminative questioning, not other aspects of dialog. Exploring other aspects of dialog (e.g., clarification questions) is interesting future work. That said, our model does exhibit continuity in the dialog. For instance, we visualize the dialog and find our Q-bot asks "is the woman alone?" followed by "what is she holding in her hand ?". The hand refers to the hand of the woman from the earlier question.

[R2] Random distractor images may be easier than a more systematic selection. We agree, but harder pools would make the task harder for all approaches. Our focus is on the relative performance of methods rather than the absolute performance. Note that we tried a harder pool by selecting visually similar images and the models' trends are similar, though overall performance was worse.

[R2] "Longer dialogs achieve better accuracy" is not supported clearly. Due to page limit, we show the task performance over rounds of dialog in Figure 8 of supplement. Performance generally goes up for Stage 2 models (trained for multiple rounds), but it goes down for Stage 1 models (only trained for a single round). This trend is very consistent across different models.

[R2] Make it clear how the target image is selected for A-Bot. Will do. The target image is randomly selected.

[R2] Missing dataset splits. Is paper using the same split as GuessWhich? Since our task contains out-of-domain images such as AWA and CUB, we can not use the same split as GuessWhich. For COCO, we use default train split and randomly split the val split into validation (30%) and test (70%). For AWA and CUB, we use the same train/val/test splits defined by the datasets.

[R2] Please clarify how interpretability is improved. We measured interpretability using fluency metrics and human performance / qualitatives (Section 4.4). Our model is able to generate questions which are more relevant to the image pool and more fluent compared to the Typical Transfer model. Furthermore, when humans are paired with our model they perform better at the game than when paired with a baseline models. This directly demonstrates interpretability: humans are able to interpret our model's responses better.

[R3] Many implementation details are referred to as "standard" or detailed only in the supplement. This may limit the accessibility of the contribution. Thanks for this valuable perspective. If granted the additional page provided to accepted papers, we will try to make more of this background available to readers.

[R4] VAEs often have the mode collapse problem where the latent structure is ignored. Did it happen here? We did not observe this in our experiments. Further, we find the latent space to be fairly well-behaved when interpolating. For example, interpolating in the latent space from "how many beds?" to "where is he looking?" yields this result after removing the replicated questions: *how many beds? - how many cats? - how many dogs? - where is the dog? - where is the man? - where is this man? - where is this woman? - where is this? - where is he? - where is he looking?*

[R4] Is the text generator robust to the distribution shift of the latent states? Based on the interpolation result above, we believe so. We hypothesize this is because the discrete latent representation and VAE pre-training help disentangle intent from expression by restricting information flow through z.

[R4] Typical Transfer has low accuracy in quite a few settings. Our hypothesis is that what enables Typical Transfer to achieve high performance for VQA is its ability to find patterns that "overfit" to Abot. These overfitting patterns are harder to find when the domain shifts (AWA/CUB).