We thank the reviewers for their time and feedback. In this paper, we argue that intrinsic motivation can take many forms. Inspired by humans, we propose a curiosity formulation based on multimodal association: searching for novel associations to explore. We demonstrate that audio-visual curiosity shows promising results not only on standard Atari environments but also on the realistic Habitat setting.

We highlight that the reviewers believe the paper is interesting (R1), novel (R4), "the first multimodal curiosity work" (R2), and shares "intuitive inspiration and some promising results" (R3). There are concerns with regards to the applicability of the approach and the failure cases, and R3 has concerns with respect to the formulation (specifically the use of error as reward). We address all of these concerns below. However, the main contribution of our paper is the introduction of multimodal curiosity. Independent of the specific formulation, this should be of great interest to the NeurIPS community and pave the way for exploiting the richness of data for better performance.

R1, R3: Formulation (R3: error as reward, R1: association is as hard, R1,3: couch-potato issue): Error as reward is not necessarily bad; the question is what error. Our approach is not to predict audio given visual features (or vice-versa). Instead, given both audio and visual features, we classify whether they are aligned or not. We highlight that this association classification error (i.e. association novelty) is fundamentally different from the prediction error based formulation typically used. In particular, this has two implications:

(a) In the prior visual curiosity framework, the model predicts the future frame in high-dimensional space. We argue that compared to higher-dimensional prediction, our classification formulation is less susceptible to the issues mentioned. For example, if pressing a button produces three distinct sounds, we could learn to classify all of these as associated, while an agent using future prediction error would always be curious.

(b) With error as reward, as Schmidhuber points out, "the problem is that in non-deterministic environments the controller will focus on parts of the environmental dynamics which are inherently unpredictable." On the other hand, in our case, the discriminator focuses on deterministic aspects to solve the alignment classification problem. This effectively helps ignore stochasticity! Therefore, while our approach would not overcome a purely random environment, it would mitigate the couch-potato problem. In contrast, future prediction-based curiosity is attracted to randomness. Our noise ablation, which adds noise to *both* the audio and the visual features, provides some insight into this.

R1, R4: Applicability/Generality: This paper has taken a step forward in terms of real-scenario experiments compared to prior work, which uses Atari as a standard benchmark. We extend to Habitat, which has many characteristics of the real world: realistic audio and visual modalities, generated from physical processes, no clear separable sound effects, and nontrivial associations (i.e. not one-to-one object-to-sound correspondence). We highlight that our method can perform without direct visibility, in the presence of background noise, and with more than one audio source:

(a) **Visibility**: Direct visibility of an object is not required for association; context should be sufficient. For example, hearing a microwave sound in a kitchen would be positively associated even if the microwave is out of view.

(b) **Background Noise**: We acknowledge that background noise could be an issue, but it would be so when there is only background or random noise (as discussed in the above formulation response). If there are foreground audio and visual signals, we can still learn associations in the presence of noise.

(c) **Multiple Audio Sources**: Multiple objects also render visual prediction hard. It requires object segmentation and modeling relationships. Similarly, multiple audio sources would require segmentation.

R1: Issue viewing supplementary material: Thank you for bringing this to our attention. We cannot change the format at this time, but you should be able to unzip it with `jar xvf supplementary material.zip`.

R2: Which baseline from Burda et al.: Our Burda et al. baseline uses random CNN features, which is stronger than pixel prediction, as you mentioned. Our method uses the same random CNN features, as shown in Figure 2.

R2: RND with audio baseline: We agree that RND would strengthen the audio in baseline ablation. In our preliminary experiments, RND with audio appears to perform similarly to the RND baseline without audio. We will include this ablation in the final paper.

R3: Definition of association: We think of association as learning shared information between modalities, i.e. the underlying physical processes that govern these different signals. We do implement this as learning alignment (are two things cotemporal), but this is not the only way to learn associations. We will modify the text to clarify this.

R3: Audio padding: Yes, the padding to 1 second is done to make each sound equal length for feature computation.

R3: Prior work on mitigating shortcomings: Thanks for pointing this out and we will add this discussion and context. Our approach, different from this body of prior work, looks at how multimodal data can mitigate these issues.

R3: Terminology, prior work, typos, citations: We are grateful to R3 for their detailed comments and will definitely incorporate this feedback into the final paper.

R4: Fixed random initialization: We use random CNN features to be similar to prior work. This is the same feature representation as used in our Burda et al. baseline.

R4: Intrinsic or extrinsic rewards: The agent only has access to intrinsic rewards, as described at L174-175. Extrinsic rewards are used only for our evaluation of exploration.