All reviewers

Thanks for valuable comments. Our code is released to clarify implementation details and facilitate future work. In Table 1 of revised paper, baselines of Gaussian, uniform and FCB are added, with CC values of 0.002, 0.003 and 0.501, respectively.

R1

1.The potential application

The potential application of HM maps is rate control (RC) of panoramic video coding, which saves ~53% bit-rate without loss of perceptual quality. Our RC proposal has been adopted by international standard of panoramic video coding (not cited due to review policy). Our RC proposal assigns bits according to pixel-wise HM maps. Compared to 2D videos, most regions of panoramic videos cannot be seen, thus having much more perceptual redundancy. In the revised paper, we will cite our proposal as application.

2.Head movement (HM) prediction V.S. eye movement (EM) prediction.

Only FoV can be seen in panoramic videos, while all regions can be viewed in 2D videos. Thus, different from 2D videos, HM prediction is the first step to predict FoV regions seen by viewers, as the objective of our paper. EM prediction is the second step to detect regions-of-interest clearly seen by viewers, still working in progress. Our RC application only uses HM prediction, already saving ~53% bit-rates, as most regions are not seen by viewers. We will discuss this.

3.Formal math definition of propositions

It is in the proof of supplementary, and will be presented in corresponding main text.

4.Distribution "subjects" in Eq. 5 & derivation of Eq. 6 for supplementary

Consider p\_n(x\_t,y\_t) denotes the probability of n-th subject views a specific HM position. Since we measure how likely all human subjects view this position, we model expectation of this probability for all human subjects (Eq. 5). Eq. 6 is derived from Eq. 3,4&5 and expectation. The derivation will be added.

5.Proof of reward functions

Another supplementary material will be added to prove Eq. 2&3. Specifically, consider we have groundtruth HM data at frame t. It can be decomposed into {(x\_n,y\_n),alpha\_n}, where (x\_n,y\_n) and alpha\_n are groundtruth HM position and direction for the n-th subject. Our goal is to estimate p(alpha|(x,y)), where (x,y) and alpha are HM position and direction for multiple subjects. We assume that probability of each pixel being the HM position decays alongside great-circle distance to (x\_n,y\_n), following Gaussian distribution. Similar decay holds for HM direction alpha. Then, we can derive Eq 2&3.

6.Missing definition of V

V is defined as expected Q under pi: E\_{a~pi}Q(a,s), and approximated by Eq. 5, similar to A3C [35].

7.Decoupling of magnitude&direction

This fastens DRL learning due to smaller action space (saving 33.2% training time), with negligible performance loss (0.0015 CC).

8.Parameters optimized on test set?

It is a typo. In fact, all hyperparameters were tuned on training set.

9.Dataset in [21]

It was not released till NIPS deadline.

R2

1. Why CNN+LSTM

HM is not only determined by current FoV (modeled by CNN), also related to previous experienced FoVs (modeled by LSTM). Similar CNN+LSTM cascade was adopted in DRL [31].

[2.HM](http://2.hm/) magnitude and direction

We will discuss decomposition of HM scan-path to magnitude and direction in main text, rather than supplementary and Fig.5.

3.Different workflows

Despite existing consistency, HM positions may be still diverse across different subjects. By running multiple workflows, diversity of HM positions can be sampled based on learnt policy (known as epsilon-greedy). Consequently, there are >1 HM positions to generate HM map, discussed in proposition 2.

4.Proof of proposition 1

We will change it to method description, instead of proof.

5.Initial location of observers

Initial location of observers is same i.e., front center, default setting in most panoramic video players.

6.Converging to certain location after some exploration?

In some cases, the subjects/agents may converge to some certain locations (e.g., one static object of interest). But sometimes they may not converge (e.g., >1 object of interest). It depends on video content.

7.Scenes of anisotropic 3D nature

We try to include diverse scenes in our database, classified in Table 1. Yes, it includes anisotropic 3D nature, like Terminator.

R3

1.Model details and notations in Fig. 6

Yes, the model is not well introduced as Fig 6 lacks annotations to mark variables. We will replace all terms in Fig.6 by notations, i.e. , Action will be replaced by \hat{nu}\_t and \hat{alpha}\_t, denoted in #197 and #202 respectively; Observation will be replaced by o\_t, defined in #192. HM speed is a typo, which should be HM magnitude, and it will be replaced by \hat{nu}\_t, denoted in #192. Fig. 6 corresponds to #192-210 in main text, and an algorithm table will be added.

2.Why predicting both \pi and V? How is V defined?

As we apply Advantage Actor-Critic method (see [35]), both pi and V are predicted. Specifically, pi is trained to maximize advantage, the term after multiplication in Eq. 6. Advantage is based on V, i.e., expected Q under pi: E\_{a~pi}Q(a,s). V is approximated by Eq. 5, in which expectation is obtained upon accumulated gradient across N workflows and reward function of Eq. 3 is used. In the revised paper, we will clarify this.

3.Details of DRL network in Fig. 6

The architecture of CNN+LSTM is briefly presented in Fig. 6, e.g., 21\*21\*32 means 32 kernels with size of 21\*21. We will add legend for DL architecture in Fig. 6. Yes, the entire network is trained from scratch.

4.Finding 2.

Finding 2 compares standard deviation and mean of HM magnitude when they are within a certain range (described in #131), which verifies the consistence when content is similar. This will move to Finding 2 to avoid the confusion.

5.t\_{max} setting

t\_{max} is determined by frame length of each processed video.

6.Using DRAW

DRAW is run, and its CC results are 0.29 (without FCB) and 0.31 (with FCB), lower than ours.