We thank all the reviewers for the valuable comments/suggestions. We summarize our responses as follows:

[Reviewer 2] *Q1: Extensive post-processing.* We mentioned on page 2 that extensive post-processing is needed in reference paper [10], while our model addressed this issue by using a 3-level hierarchy to group semantically similar patterns automatically. We compared with MAR on eye-movement data (page 8), and our model outperforms MAR.

Q2: Phase-specific words are not well defined. The model assigns a topic to each word. In Table 1, we use different colors to highlight informative words from three topics (please refer to Section 5.2). A word is highlighted if its weight is high in the assigned topic. The second 'like' in Table 1 part 1 is not highlighted because the criterion is not satisfied.

Q3: Eye movement patterns on different diseases and multiple representations of a disease. Both Experiments I and II contain images of different types of dermatological diseases. Experiment II also contains multiple images for the same type of disease. Eye movement patterns usually match the disease morphology, and thus the patterns are usually similar for the same disease. For instance, the skin disease *Halo Nevus* is usually characterized by a single discrete lesion (morphology), and the concentration pattern dominates eye movement patterns for all experts who viewed the image. We will add illustrative examples to the revised paper.

Q4: Experts not following a scheme. In this work, we study knowledge data from experts who are trained professionals. They analyze the images in systematic process, and their verbal descriptions usually follow certain schemes. A future direction is to make the model more robust to cases where careless practitioners do not follow such schemes.

Q5: Topics 29,30,32 are not included in Table 4. Table 4 matches Figure 4 (please refer to Table 4's caption), which shows 28 topics inferred from Experiment I. The 32 topics are from Experiment II.

Q6: How the controlled experimental environment was. Please refer to Section 3 for the experimental environment.

[Reviewer 3] *Q1: The use of saccades.* We use saccade features mainly for interpretation: 1) We draw lines for saccades to visualize eye movement patterns; 2) We interpret patterns as concentration/switching/cluttering using fixation duration and saccade amplitude. For model training, we only use the features of fixations.

Q2: Number of reasoning phases. The classification accuracy of disease morphology for two experiments are 81.2 and 79.8 (2 phases), 78.5 and 77.1 (4 phases). For diagnostic correctness, the results are 70.9 and 68.5 (2 phases), 69.0 and 68.2 (4 phases). According to the results, the optimal number of phases is 3 (2-phase is the second-best).

Q3: Why Lemma 1 and 2 are true. Detailed balance guarantees the convergence of the MCMC sampler (please refer to reference paper [29]), and a general form of acceptance rate is given by Eq 12. Eq 14 is extended from Eq 12 for the switch proposals. Eqs 13 and 15 show how each component in the acceptance rates are calculated and are discussed in Appendix B.5. We will add more details about the proof in the revised paper.

Q4: Mutual exclusivity of the proposals. We would like to clarify that the mutual exclusivity is not used to prove the lemmas. Mutual exclusivity ensures that all three proposals are important for speeding up MC sampling, because the effect of one proposal cannot be substituted by the effect of a combination of other proposals.

Q5: Relation to prior work. We agree with the reviewer that for eye movements, removing the hierarchy and nonparametric prior yields MAR, and for narrations, removing the hierarchy yields LDA. We will add discussions.

Q6: MN prior. We agree that the prior is essentially multivariate normal (MVN) on each column. Since the posterior follows a matrix normal distribution (MN), we write the prior in the form of MN to be consistent.

[Reviewer 4] *Q1: More examples of the topics.* We will add interpretations of the words and topics in Table 4. For instance, words in Topic 1 mainly correspond to disease morphology, including color ('brown' and 'dark'), location ('area' and 'center'); words in Topic 2 correspond to patients and body location ('patient', 'nodule', 'foot', 'dorsal').

Q2: Justification of baselines. The baselines are closely related to our model in terms of pattern discovery (auto-regression for eye movements and topic models for narrations). Comparison with those baselines illustrates the benefits of our model: using a three-level hierarchy to leverage cross-modal interactions and improve latent pattern discovery.

Q3: A careless practitioner could misuse the model. Please refer to Reviewer 2 Q4.

[Reviewer 5] *Q1: Discussion of scalability.* We would like to clarify that our model is designed for knowledge data, which is usually on a small scale due to the cost of data collection from experts. Meanwhile, it is critical to incorporate domain knowledge for pattern discovery given the complex nature of these data. We collected 807 and 720 valid eye movement/narration sequences for the two experiments, which are in a medium scale for similar domains. The design of using phases, topics, and eye movement patterns effectively incorporate prior knowledge about the nature of data (i.e., there are 3 phases; each phase has its unique topic distribution) to discover useful/interpretable knowledge patterns.

Q2: Terms in Section 4.2. 'observations of eye movements' are the recorded eye-gaze locations; 'main patterns' and 'subpatterns' are inferred latent patterns that capture high- and low-level characteristics of eye movements.

[Reviewer 7] *Q1: Presentation & Code.* We will improve the presentation to highlight the main contribution. We will add the Java code for recovering the words from indices and visualization and improve the documentation. We plan to make the code public via Github if the paper is accepted.

Q2: Terminologies/typos. IRB is Institutional Review Board (similar to ERB). We will correct other typos as suggested.