We thank all the reviewers for their insightful and constructive comments, and answer their questions below.

To Reviewer #1

(a) *line 223-224*: Yes, conditional posteriors are used here. We use q_1 if $w_{ij} = 1$ and q_0 otherwise.

(b) *Format squeezed too much, Fig. 2 difficult to decipher*: We will update Fig. 2 to improve clarity. We will also make edits to highlight the key contributions and move less relevant details to the Appendix.

(c) *What's PWA and clarify VHE's gain over PWA:* PWA refers to WANE with phrase-by-word alignment for textual feature extraction. It is a discriminative model (no prior of any sort) while our VHE is a generative solution. Generative baselines without homophilic priors are naive-VAE and VGAE. Tables 2 and 3 contain the ablation study requested by the reviewer, which decomposes the gains into individual contributions. PWA improves over WANE (prior SOTA), showing the proposed phrase-by-word alignment (a side contribution) delivers better performance. VHE's gain over PWA is more apparent on vertices with fewer connections (see Fig. 3), which demonstrates VHE's robustness and effectiveness. This also bears practical significance because low-degree vertices are what existing models struggle with.

(d) *Whether modeling of unknown links brings meaningful differences in experiments:* This corresponds to the ablation study provided in line 332-336 and Fig. 4(c). We have a hyper-parameter α to control the strength of uncertainty and observe that a proper choice of α (0.4) achieves the best results.

(e) *Limitations and prospects*: While achieving significant performance gains, the current setup of VHE only encapsulates pairwise structural information in the prior. The integration of higher-order topological information is an interesting topic, and we leave it for future investigation.

To Reviewer #2 We appreciate reviewer's acknowledgement of our novelty and constructive suggestions provided.

(a) *Contribution of phrase-to-word alignment:* While the key contribution of this work is the VHE model, our phrase-to-word alignment module also demonstrates significant performance gains over existing SOTA, which qualifies it as a side contribution. While several similar sequence-to-word attention mechanisms have been considered in other NLP tasks, the application in a network embedding context is novel.

(b) *What's "without loss of generality" in Line 77*: We mean the techniques developed can be similarly applied to directed graphs. We will clarify this in our revision.

(c) *Clarify Line 190, Line 183-195*: K_r/K_c are fixed hyper-parameters shown in Line 525-526. We will revise Line 183-195 to improve clarity.

(d) *What's PWA*: PWA refers to WANE with the proposed Phrase-by-Word Alignment. See our reply (c) to Reivewer #1 for additional details.

(e) *Improving Table 2*: Thanks for the suggestions. We will categorize the methods in Table 2 into four groups, namely topology-only baselines, topology+content baselines, generative baselines and proposed models. Table 2 will be revised accordingly, and acronyms will be clearly defined.

(f) *Response to Improvements*: We agree that the current VHE implementation fails to capture higher-order information and does not account for global topology. Extensions to these directions are interesting topics, which we are actively exploring. To note, we experimentally found that for textual network applications a fully generative solution encodes too much nuisance information, which is often detrimental to the performance, thus pooling is applied.

(g) We will fix all the grammar and formatting issues pointed out by the reviewer.

To Reviewer #3 We thank the reviewer for the positive reviews. The remarks raised are addressed below.

(a) *Why $\boldsymbol{H}$ encodes connectivity information?* The use of structural embedding $\boldsymbol{H}$ is motivated from Node2Vec [17], where it is assumed vertex-based topological profile (i.e., structure) can be encoded by a learnable vector representation. To verify $\boldsymbol{H}$ indeed captures structural information, we carried out an ablation study and summarized the results in Figure 4(d). It is clear that the use of structural embedding $\boldsymbol{H}$ improves over models that only use text information when predicting network topology. We will further clarify this.

Table 1: Computational cost for VHE.

Dataset	Train (s/epoch)	Inference (s)
Cora	2.8	45.6
HepTh	1.6	17.2
Zhihu	17.8	500

(b) *Actual computational cost.* The computational costs are charted in Table 1. This confirms VHE is very efficient in practice, and the significant performance gain fully justifies the mild increase in computation time comparing to existing SOTA. A more comprehensive discussion will be added to our revision.

(c) *Clarifications.* We will clarify that textual attributes are still available for missing vertices. (Line 227) We use 50 Monte Carlo samples to reduce the computational cost for global network embedding with each vertex.