We would like to thank all the reviewers for your helpful comments and suggestions. We appreciate your positive comments on our work: "nice structure and readability", "well-written" and "reasonable design". Hereafter, we first provide two responses to the common concerns raised by the reviewers, and then reply each reviewer, respectively.

Common Response 1: Complexity analysis of our SCAN model. As shown in Appendix A.3, the layer-wise propagation rule for both the encoder and decoder networks is the main time cost of our algorithm, while the two-layer GCN network has the highest computational complexity in the computational propagation flow. The time complexity of the two-layer GCN network for one epoch boils down to 2 sparse-dense-matrix multiplications for a cost of $O(|\tilde{\mathbf{A}}^{+}|(H+D+M+N))$, where $|\tilde{\mathbf{A}}^{+}|$ denotes the number of nonzero entires in the Laplacian matrix, H is the dimension of the first hidden layer, D is the dimension of latent embeddings, $M+N$ is the dimension of node features. Empirically, our SCAN costs around 57s and 290s per 10 epochs on the Flickr and Pubmed datasets, respectively, for training on an Inter i7 3.60GHz CPU computer.

Common Response 2: Representation updates on Section 3.2. Considering the comments of Reviewer #2 and Reviewer #3, we now revise $\mathcal{O} = (\mathcal{X}, \mathcal{Y}, \mathcal{R})$ to be $\mathcal{O} = ((\mathcal{X}^g, \mathcal{Y}^g) \times (\mathcal{X}^h, \mathcal{Y}^h) \times \mathcal{R}^{gh}), g, h \in \{1, \cdots, T\}$, which is the generalized heterogeneous data in our paper, where the relationship type $\mathcal{R}^{gh}$ is depend on the entity types g and h. In our task, we have two entity types and two relationship types, which can be represented as $\mathcal{O}_{AN} = ((\mathcal{X}^1, \mathcal{Y}^1) \times (\mathcal{X}^1, \mathcal{Y}^1) \times \mathcal{R}^{11}, (\mathcal{X}^1, \mathcal{Y}^1) \times (\mathcal{X}^2) \times \mathcal{R}^{12})$. We will revise the corresponding representations in Section 3.2 based on this formulation and focus more on the specific heterogeneous data type in our task (i.e. $\mathcal{O}_{AN}$). For instance, Eq (4) will be revised to $p_\theta(\mathcal{O}_{ij}, \mathcal{Z}_{ij}) = p_\theta(r_{ij}^{gh} \mid \mathcal{Z}_{ij}, \mathbf{Y}^l)p(\mathbf{Y}^l)\prod_{\mathbf{z}\in\mathcal{Z}_{ij}} p(\mathbf{z})$. We will further provide a Section to discuss the potential extension of our model to handle other types of heterogeneous data, such as bipartite graph in recommender system, multi-relationship in knowledge graph and multi-type of entities; and clarify the difference on the variation and the capability of our model when handling these type of heterogeneous data.

Response to Reviewer #1

- Thank you for reviewing our paper and noticing the typos and improper notations in our paper. We have double checked the whole paper and corrected all the typos we can find.
- Regarding the performance on varying training set sizes, we have conducted experiments. Here we only show the accuracy result on the BlogCatalog dataset, and detailed analysis will be provided in the Appendix of our final version.

% training	85%	80%	75%	70%	65%	60%	55%	50%
SCVA_DIS	.845	.838	.833	.828	.810	.792	.780	.778

- Regarding the complexity analysis of our model, please see the response in **Common Response 1**.

Response to Reviewer #2

- Thank you for reviewing our paper and noticing the typos and confusing definition in our paper. We have double checked the whole paper, corrected all the typos we can find and enlarged Figure 4 to make it more legible.
- Thank you for the suggestions on Section 3.2. Please see the response in **Common Response 2**.
- Eq (4) is a generalized form of the factorization of the joint probability with two entities, their labels and their relations, where we assume that the labels of two entities are not explicitly given in current form. We now modified Eq (4) to make it more clear (See **Common Responose 2**), where the full derivation will be in final version. The specific factorization forms on different cases can be found in the Appendix (Please see Eq (12), Eq (16), Eq (20), Eq (24) and Eq (28)).
- We now provide the complexity analysis of our model; please see the response in **Common Response 1**. We will also provide the statistics information of the 3 datasets and report the the runtime of SCAN with each setting in the Appendix of our final version.
- We tuned the latent dimension D in {16, 32, 64, 128}, and chosen D with a best performance on the validate sets in these tasks. In our paper, $D = 64$ was a default setting, with which the best performance is achieved in all the datasets.
- In the attribute inference task, our SCAN did make use of the label information to obtain their embeddings. We indeed have made comparison with state-of-the-art unsupervised methods (such as CAN) in terms of node classification task, and our model shows better performance. For fair comparison we only report the result on semi-supervised task.
- Thanks for the good suggestions. We will move the performance result over different label ratio to the main text, and seek to evaluate our proposed model on more topologically diverse real datasets and report result in the final version.

Response to Reviewer #3

- Thank for the good comment on "little over-claimed". Please see the response in **Common Response 2**.
- The title of Section 3.2 will be revised to "Semi-supervised Learning for Heterogeneous Data".
- Thank you for noticing the typos. In Eq (5), $\mathbf{Y}^l$ in q_ϕ should be omitted, and log should be added to $p(z)p(y)$.
- There is no sum on p_θ, since $\mathcal{O}_{ij}$ is an atomic data point. Please see the revised form of Eq (4) in **Common Response 2**, where the full factorization will given in the appendix of the final version. q_ϕ can be factorized in different forms according to different cases (Please see Eq (13), Eq (17), Eq (21), Eq (25) and Eq (29) in the appendix of the paper).