We appreciate all reviewers' valuable comments, and greatly encouraged by the positive comments, e.g. the problem studied in this paper is interesting and practical; this is an interesting method that is substantially novel; the approach looks sensible and the derivation appears sound; the experimental validation is solid. We shall address the main concerns point by point as follows.

To Reviewer 1 & 2 on missing recent related methods. We had conducted extra experiments on more recent related methods as listed in the table below, but initially we did not include them in our submission considering that it would be too messy to list the results of all methods, which was also not allowed due to the space limit. Thus, we chose to report those were 1) representative that served as baselines for most of the related works 2) as interpretable as our method, although we compared to some advanced deep models to show the effectiveness of our method, and 3) scalable to the scale of datasets in our experiments. Of course, we do agree that our submission can be more persuasive by including more results of the SOTA methods. We would include these results in supplementary materials of the final version.

To Reviewer 1: Q1 - The improvement of your method is not impressive. As mentioned with references in line 224 of our paper, an improvement of 0.001 on Logloss is significant on the two datasets Avazu and Criteo, considering the size of datasets and business value of related tasks. Our model improves Logloss by 0.002 and produces more interpretable results compared to the second-best method which is based on neural networks. Thus we claim the results to be significantly better, although more #params are used. **Q2 - I want to know more about the datasets, which is different about traditional dataset.** The datasets are two publicly available advertisement click prediction datasets with anonymous features and columns. Some columns contain site_id and ad_id so the dimensionality is very large. We would include more details in supplementary materials of the final version.

To Reviewer 2: Q1 - The proposed method lacks technique contributions. We admit that our idea is simple and straightforward, but based on our knowledge such learning strategy has not been explored for multi-field categorical data. **Q2 - The generalization error bound seems tight.** Our generalization bound is tight compared to many SOTA papers regarding over-parametrised models. For example, related bound on Criteo dataset is tighter than initialisation-based bounds calculated with [4] as we observed in our experiments. **Q3 - More real-world data should be considered.** We had conducted more experiments on datasets such as MovieLens and Frappe. Our model consistently outperformed other baselines, but given the page limitation, we could only report the results on two larger datasets Avazu and Criteo.

To Reviewer 3: Thanks for your suggestions and we will revise our paper accordingly. **Q1 - Does the improvement come from the specific model structure or the aggregation function F.** We set the aggregation function F to a simple sum function in our paper so that we can mainly credit the performance gain to the specific model structure. **Q2 - one should use different values of r.** This is exactly what we did in the experiments. We chose different ranks for each field in a log scale regarding the cardinalities of each field as mentioned in line 254 of our paper. **Q3 - How could the method be extended to handle both categorical and continuous features?** 1) we could transform the continuous features to categorical features by log transformation as mentioned in line 209 of our paper 2) or we could conduct field-wise learning only on categorical fields, and directly use continuous features in each field-focused model. **Q4 - is building a model for every individual field really the way to go when the number of fields is large?** We may have to selectively choose some of the fields for field-wise learning on datasets with more fields and features.

To Reviewer 4: Q1 - The connection to other multi-view methods. We agree with you that in the higher-level concept the method can be related to multiview learning. Here we focus more on the special structure of categorical variables, which is discussed less in general multiview learning method. We would add a short discussion on the connection to multi-view methods. **Q2 - For reducing the number of parameters, maybe a sparsity inducing norm rather than an L_2 norm on W would work.** Thanks for the suggestion and we may investigate it in our future work. We did not pose L2 norm constraint on W but on column average of W to promote the generalisation ability. To reduce the parameters, we decomposed W into two much smaller low-rank matrices U and V.

	Avazu				Criteo			
Method	Logloss	AUC	Time	#params	Logloss	AUC	Time	#params
RaFM [1]	0.3774	0.7862	20h58m	86.53M	0.4417	0.8104	10h22m	29.26M
IFM [2]	0.3746	0.7885	30m	82.74M	0.4403	0.8118	3h20m	16.07M
AFN [3]	0.3768	0.7857	3h26m	107.09M	0.4406	0.8118	10h14m	16.71M

[1] Xiaoshuang Chen, et al. "RaFM: Rank-Aware Factorization Machines." ICML 2019.

[2] Yantao Yu, et al. "An Input-aware Factorization Machine for Sparse Prediction." IJCAI 2019.

[3] Weiyu Cheng, et al. "Adaptive Factorization Network: Learning Adaptive-Order Feature Interactions." AAAI 2020.

[4] Behnam Neyshabur, et al. "The role of over-parametrization in generalization of neural networks." ICLR 2019.