
From Feature Interaction to Feature Generation: A Generative Paradigm of CTR Prediction Models

Mingjia Yin^{1*} Junwei Pan² Hao Wang¹ Ximei Wang² Shangyu Zhang² Jie Jiang² Defu Lian¹
Enhong Chen¹

Abstract

Click-Through Rate (CTR) prediction models estimate the probability of users clicking on items based on feature interactions, inherently following a discriminative paradigm. However, this paradigm is prone to embedding dimensional collapse and information redundancy due to limitations of vanilla feature embeddings. This motivates us to reformulate it into a generative paradigm to generate new feature embeddings. Unlike sequential recommendation, which naturally fits a generative "next-item prediction" paradigm, it's hard to formulate CTR models into this paradigm because there are no explicit orders between features. In this paper, we propose a novel *Supervised Feature Generation* framework for CTR models, *shifting from the discriminative "feature interaction" paradigm to the generative "feature generation" paradigm*. Specifically, we predict each feature embedding based on the concatenation of all feature embeddings. Besides, this paradigm naturally accommodates a *supervised binary cross-entropy loss* to indicate whether the sample is positive or negative. The framework can reformulate nearly every existing CTR model and bring significant performance lifts. Moreover, it produces less-collapsed and redundancy-reduced feature embeddings, thereby mitigating the inherent limitations of the discriminative paradigm. The code can be found at <https://github.com/USTC-StarTeam/GE4Rec>.

*Work done while an intern at Tencent Inc. ¹State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei, China ²Tencent Inc, China. Correspondence to: Hao Wang <wanghao3@ustc.edu.cn>.

1. Introduction

Click-Through Rate (CTR) prediction models estimate the probability of users clicking on items based on feature interactions. While conventional wisdom holds that CTR models inherently follow a discriminative paradigm, they are prone to embedding dimensional collapse (Guo et al., 2024) and information redundancy (Zbontar et al., 2021), primarily due to the limitations of vanilla feature embeddings. Therefore, we explore reformulating CTR models into a generative paradigm to generate new feature embeddings.

Different from sequential recommendation models (Kang & McAuley, 2018; Rajput et al., 2023; Zhai et al., 2024; Yin et al., 2024), very limited research has focused on formulating CTR models under a generative paradigm. This is possibly due to the fact that there are no explicit partial orders among the inputs of CTR models, making it difficult to directly fit them into the popular next-token (Vaswani et al., 2017) or next-scale (Tian et al., 2024) prediction paradigm.

In auto-regressive models, as illustrated in Fig. 1a, a widely adopted approach is to consider the sequence up to position N as the *source input* and the item at position $N + 1$, *i.e.*, the next token, as the *target input*. Alternative formulations also exist; for instance, masked generative models (Fig. 1b) (Chang et al., 2022; Li et al., 2023) treat the entire input as the *source input* and randomly selected portions as the *target input*, while VAR (Fig. 1c) (Tian et al., 2024) considers coarse-scale token maps as the *source input* and fine-scale token maps as the *target input*.

CTR prediction handles the multi-field categorical data (Zhang et al., 2016; Pan et al., 2018), where there are usually no explicit partial orders between input features. To this end, we propose to treat each feature as the *target input*, and build an *Encoder-Decoder* network upon the whole input features, *i.e.*, the *source input*, to predict it. Such a method can be regarded as a "feature generation paradigm", which models the $P(X)$ on the unordered multi-field categorical data.

Specifically, the encoder employs a non-linear activated MLP upon the concatenation of all feature embeddings, generating a new representation for each feature. Then we use

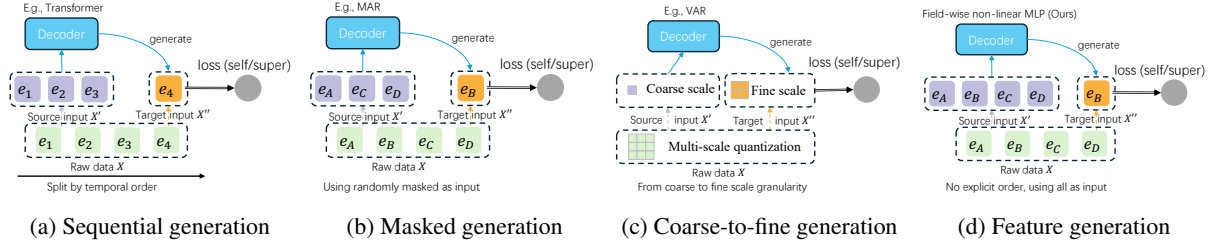


Figure 1. Different generative paradigms. Existing autoregressive generative models can be unified under a common framework. Specifically, these models construct an *encoder* based on the *source input*, generate an output embedding, and utilize it to predict the *target input*. The core of paradigm design involves selecting appropriate *source input* and *target input* based on the intrinsic data structure.

this new representation to predict the embedding of another feature i using various feature interaction functions such as the dot product or the projected product (Sun et al., 2021; Wang et al., 2021). Such a design *avoids direct interaction (product) between vanilla ID embeddings* as done in traditional CTR models (Rendle, 2010; Guo et al., 2017; Sun et al., 2021; Wang et al., 2021), and hence prevents the embeddings from dimensional collapse due to Interaction-Collapse Theory (Guo et al., 2024). Besides, the output embedding of *encoder* can be treated as a new representation of each feature that is tailored to different samples. These new representations are more decorrelated with the vanilla representation, fulfilling the redundancy reduction principle (Zbontar et al., 2021).

Conventionally, generative paradigms are accompanied by a self-supervised loss. In the next-prediction paradigm in sequential modeling, the label of the next item is usually a self-supervised signal, that is, whether the next item is the ground-truth one, or just a random one. However, in feature interaction recommendation, it is unnecessary since natural supervised signals already exist. Therefore, we adopt a supervised loss with the proposed feature generation paradigm rather than the widely adopted self-supervised loss in the next-prediction paradigm. The feature generation, together with the supervised loss, leads to a novel *Supervised Feature Generation* framework for CTR prediction.

This framework can reformulate nearly every existing feature interaction model, ranging from FM to DeepFM, xDeepFM, and DCN V2. Comprehensive experiments demonstrate that this new framework significantly improves performance, achieving an average of 0.272% AUC lift and 0.435% Logloss reduction, while incurring only a marginal increase in computational overhead—an average increase of 3.14% in computation time and 1.45% in GPU memory consumption. It can produce feature embeddings with reduced collapse and redundancy compared to raw ID embeddings. Additionally, we conduct extensive ablation studies to validate the framework design. We successfully deploy it to one of the world’s largest advertising platforms for click prediction, with a 2.68% GMV lift on a primary scenario,

leading to one of the largest model improvements in 2024.

2. Method

In this section, we first define the problem and introduce traditional discriminative CTR paradigms and their limitations in Sec. 2.1. Subsequently, in Sec. 2.2, we propose a novel generative feature generation framework based on our findings. Finally, Sec. 2.3 presents the detailed implementation of the proposed supervised feature generation framework.

2.1. Preliminary: CTR prediction in a discriminative paradigm

Problem definition. CTR prediction aims to predict the probability that users will click on items based on multiple features. The problem can be formally defined using features $\mathcal{X} \in \{0, 1\}^M$ and a label set $\mathcal{Y} \in \{0, 1\}$, indicating whether users click the target item. Typically, $\mathcal{X}$ consists of multiple feature categories, including user, item, and context features. Suppose there are N different feature categories, each category can be represented by a distinct feature field F_i of cardinality D_i , where D_i measures the number of unique features in each field.

Formulation. In general, CTR models learn a function Φ mapping $\mathcal{X}$ to $\mathcal{Y}$, which models the distribution $P(\mathcal{Y}|\mathcal{X})$. This aligns with the definition of discriminative models, thereby existing CTR models can be naturally formalized as the following discriminative form:

$$\mathcal{L}(y_{\text{super}}, \overbrace{h_{\text{classifier}}(g_{\text{inter}}(\{v_i\}))}^{\text{classify}}). \quad (1)$$

feature interaction

$v_i \in \mathbb{R}^K$ is obtained from the embedding lookup table $V_i \in \mathbb{R}^{D_i \times K}$ corresponding to field F_i , where K denotes the embedding dimension. g_{inter} is the feature interaction module exploring high-order correlations among feature fields, usually achieved through hadamard product between feature embeddings. $h_{\text{classifier}}$, implemented with a pooling function or an MLP, finally maps the interacted embedding to a predicted label. $y_{\text{super}} \in \mathcal{Y}$ is the corresponding label

used to calculate the binary cross entropy loss $\mathcal{L}$. Using the classic DCN V2 (Wang et al., 2021) model as an example, it can be formalized as:

$$\mathcal{L}(y_{\text{super}}, \text{DNN}(\sum_{l=1}^L \sum_{i=1}^N \sum_{j=1}^N \mathbf{v}_j^{(0)} \odot \mathbf{v}_i^{(l)} \mathbf{M}_{F(i) \rightarrow F(j)}^{(l)})), \quad (2)$$

where L denotes the number of cross layers; l denotes the layer index; N denotes the total number of features; i and j denote the feature indices; $\mathbf{v}_i^{(0)}$ denotes the embedding of feature i in the embedding layer; $\mathbf{v}_j^{(l)}$ denotes the embedding of the j -th feature in the l -th layer; $\mathbf{M}_{F(i) \rightarrow F(j)}^{(l)}$ denotes the projection matrix between the $F(i)$ and $F(j)$ field pair in the l -th layer; and $F(i)$ and $F(j)$ denote the fields of the feature i and j , respectively.

Discussions on limitations of discriminative paradigms. Despite the achieved achievements, these methods still fall into a discriminative paradigm, and very limited research has focus on formulating them under a generative paradigm. Such a discriminative paradigm will encounter the following issues:

1. **Dimensional Collapse caused by raw ID embedding interaction.** The embeddings of some fields may only span a low-dimensional space due to various reasons, such as the low cardinality of this field. For example, the embeddings of the gender field with values of Male, Female, and Unknown can only span a 3-dimensional space. According to the Interaction-Collapse-Theory (Guo et al., 2024), the interactions with these low-dimensional field embeddings may lead to the dimensional collapse (Jing et al., 2021) of the embeddings of the other fields, thereby limiting their information abundance.
2. **Limitation to learn data distribution.** Discriminative paradigms learn the distribution $P(\mathcal{Y} | \mathcal{X})$ while overlooking $P(\mathcal{X})$, focusing solely on establishing a feasible decision boundary for classification (Harshvardhan et al., 2020; Oussidi & Elhassouny, 2018). However, given the inherent limitations of raw ID embeddings—primarily caused by data distribution—the lack of consideration for data distribution makes it challenging to learn an effective decision boundary.
3. **Information redundancy.** Redundancy-reduction principle (Barlow et al., 1961) has been fruitful in different application domains (Barlow, 2001; Grill et al., 2020; Zbontar et al., 2021). This principle necessitates the minimization of information redundancy between the two views, that is, their mutual correlation. We have empirically verified this principle in Sec. 3.3.2, and find that models with redundancy reduced interacted embeddings achieve better recommendation performance (Fig. 4).

But interacted raw ID embeddings with raw ID embeddings, even though after some elaborately designed transformation (e.g., DCN V2 (Wang et al., 2021)), still exhibit a strong tendency towards containing homogeneous information, i.e., information redundancy. This limits the knowledge efficiency of feature interactions.

These limitations call for a rethinking of feature interaction models in the context of paradigm designs.

2.2. CTR prediction in a generative paradigm

Autoregressive generative paradigms. Generative models focus on learning the joint distribution $P(\mathcal{X}, \mathcal{Y})$. Among generative paradigms, autoregressive approaches remain the de facto standard, predicting the next item in a sequence based on preceding inputs. Conventional wisdom posits that autoregressive models inherently require explicit partial data ordering, as demonstrated by the success of sequential recommendation systems (Zhai et al., 2024; Rajput et al., 2023). However, advancements in autoregressive models within computer vision—notably the next-scale prediction paradigm (Tian et al., 2024)—have shifted the focus toward designing generative frameworks that align with inherent data characteristics, rather than enforcing autoregressive structures on non-sequential data. This implies that explicit data ordering is not fundamentally necessary for autoregressive generation. Instead, the essence lies in modeling appropriate intrinsic data structures.

Reformulation. We re-conceptualize the generative models in feature interaction models by shifting from the "next-item prediction" to the "feature generation" paradigm. To begin with, we can integrate previous autoregressive generative models into a unified framework as illustrated in Fig. 1: Specifically, it builds a *encoder* upon the *source input*, gets an output embedding, and uses it to predict the *target input*. It can be formally defined as:

$$\mathcal{L}(y_{\text{super}}, \overbrace{h_{\text{classifier}}(\Phi)}^{\text{classify}}), \quad (3)$$

where Φ is the feature generation framework:

$$\Phi = f_{\text{encoder}}^j(\{\mathbf{v}_i\}_{\text{source}}) \odot f_{\text{transform}}\{\mathbf{v}_i\}_{\text{target}}, \quad (4)$$

where some feature embeddings $\{\mathbf{v}_i\}_{\text{source}}$ are fed into a f_{encoder}^j to construct new embeddings, which will further be used to perform generation along with the transformed target feature embeddings $\{\mathbf{v}_i\}_{\text{target}}$. The hadamard product operation is where feature generation happens, which will be further processed with different pooling function.

By transcending data sequentiality constraints, we can incorporate established feature interaction models into this generative framework. For CTR data defined in Section 2.1, explicit correlation data between feature fields remain absent.

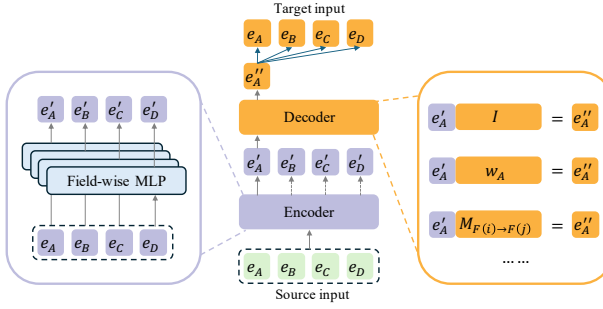


Figure 2. The feature generation framework builds a *encoder* based on the *source input*, generates an output embedding, and utilizes it to predict the transformed *target input*. Feature generation occurs through the Hadamard product operation, followed by further processing with a pooling function. For multi-layer generation, representations produced by the pooling function will be used as the new "feature ID embeddings" in the next-layer generation. Specifically, the *encoder* is implemented as a field-wise single-layer non-linear MLP.

Nevertheless, there are fundamental connections among feature fields.

Based on the above analysis, when constructing features, we employ all raw embeddings as *source input* to model implicit inter-field relationships. Following conventional generative paradigms, we designate each feature as *target input*, applying the transformations from existing discriminative model (e.g., DCN V2’s transformation matrix). The *encoder* architecture remains flexible, requiring only effective mediation between *source* and *target inputs* (implementation specifics in Section 2.3).

Conventionally, generative paradigms like next-item prediction are accompanied by a self-supervised loss, while the next item may be the ground-truth one or just a random one. Unlike conventional unsupervised generative paradigms, our approach can leverage the inherently available supervisory signals to achieve robust generative learning. The paradigm overview is presented in Fig. 1d.

Then using stacked DCN V2 as an example, Equation 2 can be reformulated as:

$$\mathcal{L}(y_{\text{super}}, \text{DNN}(\sum_{l=1}^L \sum_{i,j=1}^N f_{\text{encoder}}^j([\mathbf{v}^{(l)}]) \odot \mathbf{v}_i^{(l)} \mathbf{M}_{F(i) \rightarrow F(j)}^{(l)})). \quad (5)$$

More formal reformulations can be found in Appendix A.

Discussion. Through the feature generation framework, we have shifted from the discriminative paradigm that involves direct interactions between raw ID embeddings to a generative paradigm that constructs new embeddings by an encoder network and interacts the constructed embeddings with the raw ID embeddings. It integrates all feature fields to construct tailored features for each sample, which can avoid direct interactions between raw ID embeddings, thereby mitigating the risk of dimensional collapse (verified in Sec. 3.3.1). Besides, the constructed embeddings can be easily de-correlated with raw ID embeddings, thereby reducing information redundancy (verified in Sec. 3.3.2).

2.3. Implementation

Our main focus is on designing a generative framework suitable for feature interaction models, so we keep a simple *encoder* architecture design. We instantiate the *encoder* as a field-wise single-layer non-linear MLP:

$$f_{\text{encoder}}^j([\mathbf{v}]) = \sigma([\mathbf{v}] \cdot W_{F(j)}), \quad (6)$$

where σ is a non-linear activation function, $[\mathbf{v}] \in \mathbb{R}^{NK}$ the concatenation of all feature embeddings, and $W_{F(i)} \in \mathbb{R}^{NK \times K}$ a field-wise weight matrix. All these three components, *i.e.*, the non-linear activation function σ , the feature concatenation $[\mathbf{v}]$ and the weight matrix $W_{F(i)}$ are necessary, and we’ll present the ablation study on them in Sec. 3.4. Now, we can reformulate existing CTR models within the feature generation framework, which are detailed in Fig. 2. Notably, we regard the elaborately designed transformations in existing models as the transformation to *target input*.

3. Experiments

In this section, we aim to address these research questions:

- **RQ1:** To what extent can the paradigm shift improve existing discriminative feature interaction models?
- **RQ2:** Can the generative paradigm mitigate the inherent drawbacks of raw ID embeddings in discriminative paradigms, specifically in terms of embedding dimensional collapse and information redundancy reduction?
- **RQ3:** Is the current paradigm design optimal for feature generation? What will happen if we use different *source input*, *encoder*, or *target input*?

3.1. Setup

Datasets & Evaluation protocols. In this work, we have conducted experiments based on two widely adopted large-scale datasets, namely Criteo (cri, 2014) and Avazu (ava, 2014). Dataset statistics are summarized in Appendix B.1. As for evaluation, we evaluate the recommendation performance with AUC and Logloss.

Baselines. To verify versatility of our method, we integrate it with various representative models, including explicit feature interaction models FM (Rendle, 2010), FmFM (Sun et al., 2021), CrossNetv2 (Wang et al., 2021), and DNN-based models DeepFM (Guo et al., 2017), IPNN (Qu et al., 2016), xDeepFM (Lian et al., 2018), DCN V2 (Wang et al., 2021). All experiments are based on a popular library FuxiCTR (Zhu et al., 2020; 2022). More details can be found in Appendix B.2. Besides, the computational complexity are provided in Sec. B.3.

3.2. Recommendation performance comparison between discriminative and generative paradigms (RQ1)

Offline results. We apply the feature generation framework with various recommendation models, with results presented in Tab.4. Overall, the proposed method exhibits promising effectiveness and achieves consistent performance lift across different models, achieving an average of 0.272% AUC lift and 0.435% Logloss reduction. Usually a 0.1% AUC (gAUC) lift is regarded as a huge improvement in recommendation systems (Zhu et al., 2022).

Specifically, generative paradigm on explicit feature interaction models can bring an average of 0.428% AUC lift and 0.689% Logloss reduction. Notably, DCN V2 incorporates a DNN based on CrossNet, enhancing its modeling capability. The discriminative version of DCN V2 surpasses the discriminative CrossNet by 0.157% lift in AUC and 0.235% in Logloss reduction. Surprisingly, when CrossNet is reformulated within our generative paradigm, it can even outperform the discriminative DCNv2 by 0.106% lift in AUC and 0.089% reduction in Logloss, verifying the promising potential of generative paradigms. For DNN-based models, the improvement is less pronounced. Nevertheless, even when integrated with these complex DNN-based models, the paradigm shift still brings significant enhancements, achieving an average improvement of 0.116% in AUC and 0.181% in Logloss reduction. We conclude the following result:

Result 1. By shifting from discriminative to generative paradigm, our proposed feature generation framework brings consistent performance lift on various existing feature interaction models.

Online A/B Testing. We deployed the proposed generative paradigm in one of the world’s largest advertising platforms. The production model employs Heterogeneous Experts with Multi-Embedding architecture (Guo et al., 2024; Su et al., 2024; Pan et al., 2024). We switch the IPNN expert in the production model into a generative paradigm, which models the interactions between more than five hundred user-, ad-, and context-side features. During the one-week 20% A/B testing, demonstrated promising results, achiev-

ing 2.68% GMV lift and 2.46% CTR lift on several vital scenarios, including Moments pCTR, Content and Platform pCTR, and DSP pCTR. These improvements were statistically significant according to t-tests. The proposed feature generation framework has been successfully deployed as the production model in the above-mentioned scenarios, leading to a revenue lift by hundreds of millions of dollars per year.

3.3. How does the generative paradigm work? (RQ2)

3.3.1. GENERATIVE PARADIGM MITIGATES EMBEDDING DIMENSIONAL COLLAPSE

Dimensional collapse evaluation protocols. Dimensional collapse means that the embeddings only span a low-dimensional subspace of the available representation space (Jing et al., 2021; Guo et al., 2024), which is usually measured with singular value decomposition. Specifically, we evaluate the dimensional collapse issue at the sample level. We begin by obtaining the sample embedding matrix $Z \in \mathbb{R}^{B \times K}$ using the validation dataset, where B denotes the batch size and K the dimension size (Notably, this batch-wise setting will greatly enhance the analysis efficiency, and we have verified the robustness of this setting in Appendix B.4). The covariance matrix is then derived as $C = \frac{1}{B} \sum_{i=1}^B (z_i - \bar{z})(z_i - \bar{z})^T$, with $\bar{z} = \frac{1}{B} \sum_{i=1}^B z_i$. Subsequently, we determine the singular values $S = \text{diag}(\sigma^k)$ of C via singular value decomposition (SVD) and normalize them by the maximum singular value: $S' = \text{diag}\left(\frac{\sigma^k}{\max(\sigma^k)}\right)$. Finally, we present these normalized singular values in descending order, as shown in Fig. 3.

Evaluated embeddings. We focus on the direct impact of the feature generation framework on the embedding space. Specifically, we analyze the embedding used to interact with raw ID embeddings. In the discriminative paradigm, it is the concatenation of raw ID embeddings, formally defined as $[v]$. In the generative paradigm, it is the embedding immediately constructed by the *encoder*, formally defined as $[\sigma([v]W_{F(i)})]$. We study this embedding to investigate the direct influence of the generative paradigm.

Generative paradigm mitigates dimensional collapse.

For brevity, we illustrate the singular value spectrum of the embedding space for four representative models in Fig. 3. Visualization of all models can be found in Appendix C.1. In each sub-figure, the spectrum exhibits a rapid decay. Taking Fig. 3d as an example, the singular values of DCN V2 on Criteo remain high up to index 250, with values around 1×10^{-5} . However, they drop dramatically to 1×10^{-15} at index 280, a reduction of 10^{10} times. This indicates an extreme imbalance among dimensions, i.e., only a minority of dimensions dominate the embedding space. After index 280, the singular values remain around 1×10^{-15} , essentially

Table 1. Recommendation performance of models with the DIScrimative (DIS) and GENrative (GEN) paradigm. We conduct a two-tailed T-test to calculate the statistical significance, with results presented in a form of *mean(variance)*. **Bolded values** refer to the best performance, and * means the corresponding p-values are less than 0.05.

Model		Criteo		Avazu	
		AUC↑	Logloss↓	AUC↑	Logloss↓
Explicit	FM	DIS	0.80236(9e-05)	0.44889(7e-05)	0.78877(1e-04)
		GEN	0.81108(1e-04)*	0.44077(1e-04)*	0.79260(1e-04)*
	FmFM	DIS	0.80552(3e-04)	0.44626(3e-04)	0.78990(2e-04)
		GEN	0.80992(7e-04)*	0.44258(8e-04)*	0.79266(9e-05)*
	CrossNet V2	DIS	0.81312(1e-04)	0.43918(2e-04)	0.79106(1e-04)
		GEN	0.81540(4e-05)*	0.43661(5e-05)*	0.79301(2e-04)*
DNN-based	DeepFM	DIS	0.81380(8e-05)	0.43804(6e-05)	0.79285(1e-04)
		GEN	0.81396(6e-05)*	0.43788(5e-05)*	0.79333(7e-05)*
	xDeepFM	DIS	0.81365(1e-04)	0.43819(1e-04)	0.79222(1e-04)
		GEN	0.81421(7e-05)*	0.43775(9e-05)*	0.79429(1e-04)*
	IPNN	DIS	0.81341(5e-05)	0.43850(2e-05)	0.79348(3e-04)
		GEN	0.81415(8e-05)*	0.43776(1e-04)*	0.79451(8e-05)*
	DCN V2	DIS	0.81387(6e-05)	0.43826(4e-05)	0.79282(2e-04)
		GEN	0.81472(6e-05)*	0.43713(5e-05)*	0.79342(5e-05)*

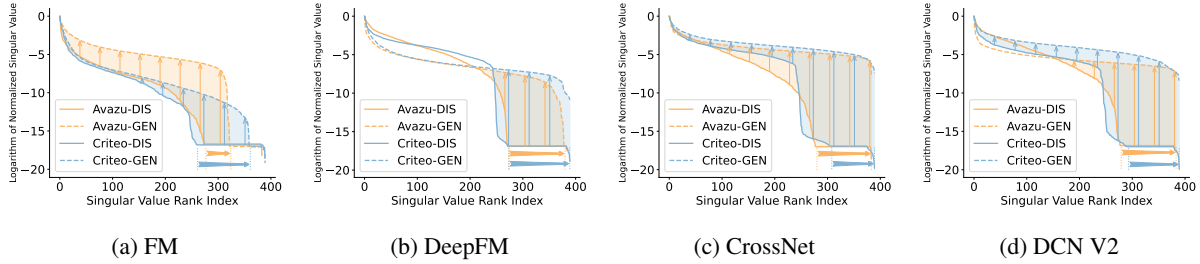


Figure 3. Normalized singular value spectrum of embeddings used to interact with raw ID embeddings. It is the concatenation of raw ID embeddings for the discriminative paradigm, while the embedding immediately constructed by the *encoder* for the generative paradigm.

zero. These singular values account for approximately 30% of the total singular values, implying that 30% of the dimensions in the embedding carry no meaningful information, which is clearly unfeasible.

These phenomena are significantly mitigated in the generative one. With the exception of FM, the singular value spectra of the other methods do not exhibit the abrupt decay mentioned earlier. Instead, they decline at a relatively slower rate, indicating a more balanced embedding space. Even for simple models like FM, our generative paradigm can increase the number of meaningful dimensions by 25%. We attribute this improvement to the integration of all feature fields when constructing embeddings using the feature generation framework. We conclude the following result:

Result 2. The generative paradigm substantially mitigates the issue of embedding dimensional collapse.

3.3.2. REDUNDANCY REDUCTION VIA GENERATIVE FEATURE LEARNING

Information redundancy evaluation protocols. According to the information redundancy reduction principle (Barlow et al., 1961; Zbontar et al., 2021), the two interacted embeddings are expected to exhibit low correlation. To quantify this, we employ the Pearson Correlation Coefficient between each dimension of the two interacted embeddings, defined as $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{s_X s_Y}$. For the discriminative paradigm, X and Y are the two interacted embeddings, while X and Y are the transformed *source input* and *target input* embeddings for the generative paradigm. s_X and s_Y denote their respective standard deviations.

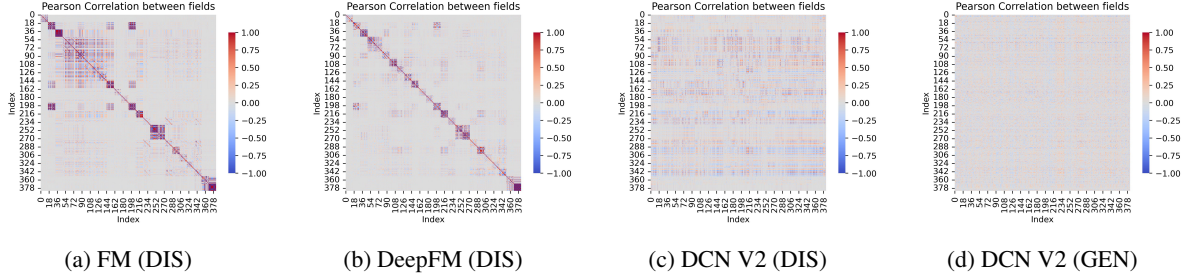


Figure 4. Pearson correlation matrix between two interacted embeddings. (a)→(b)→(c) means more complex models, which also exhibits a trend of redundancy reduction. This reveals the importance of information redundancy reduction when designing CTR models. In (d), we can find the generative DCN V2 almost produces a zero correlation matrix, perfectly aligning with the redundancy reduction principle.

Negative connection between redundancy metric and recommendation performance. We have visualized the correlation matrix of FM, DeepFM, DCN V2 in Fig. 4. More visualizations are provided in Appendix C.3. In Fig. 4a, we have derived two major observations: (1) Intra-field correlation. It forms some obvious diagonal blocks, while each block corresponds to a feature field. This means the information within a feature field is highly correlated, *i.e.*, redundant information. (2) Inter-field correlation. The index 32 - 160 forms a big diagonal block, which is exactly the correlation between field with index 3 - 10. This means these feature fields are also highly correlated, violating the redundancy reduction principle. These observations may explain the inferior performance of FM.

Then we analyze by comparing different models. DeepFM builds a parallel DNN upon FM, which greatly decreases inter-field correlation and increases recommendation performance. DCN V2 further incorporates a more advanced explicit feature interaction module based on DeepFM. In Fig. 4c, the diagonal blocks representing intra-field correlation are almost reduced, which explains the recommendation performance lift. *All these results reveal a negative connection between the redundancy metric and recommendation performance, which can guide model designing.*

Generative paradigm reduces information redundancy. Despite the transformations applied to raw ID embeddings in DCN V2, we still observe correlations in Fig. 4c. In contrast, the correlation matrix is nearly a zero matrix in Fig. 4d, indicating that the two vectors are highly de-correlated and thus adhere to the redundancy reduction principle. This demonstrates our framework’s ability to reduce information redundancy effectively. We conclude the following result:

Result 3. The feature generation framework produces embeddings highly de-correlated with raw ID embeddings, adhering to the redundancy reduction principle.

3.4. Ablation on the feature generation framework design (RQ3)

For simplicity, all ablation studies are based on DCN V2.

Ablation on the source input design. As stated in Discussion 1 of discriminative paradigms, a major limitation of them is the inherent drawbacks of raw ID embeddings, especially those of low-cardinality fields. To tackle this issue, we propose to utilize all field embeddings as *source input* for all fields. For comparison, we will investigate the following configurations to reveal the significance of our design: (a) using only the field’s own embedding as *source input* for all fields; (b) using all field embeddings as *source input* for 10 fields with the highest cardinality; and (c) using all field embeddings as *source input* for 10 fields with the lowest cardinality. Results are presented in Fig. 5a.

We can observe that using all field embeddings as *source input* outperform other settings, revealing the necessity of constructing embeddings with all features. Additionally, the results of only constructing low-cardinality fields with all features are significantly better than the high-cardinality counterparts, which corroborates our previous assertion that low-cardinality fields suffer from severe information insufficiency compared with high-cardinality fields. We conclude the following result:

Result 4. For feature generation, it is effective to use all feature fields as source input. In particular, low-cardinality field embeddings suffer from more severe issues than high-cardinality field ones, underscoring the importance of leveraging all feature information to generate these embeddings.

Ablation on the encoder design. The adopted *encoder* is a field-wise one-layer non-linear MLP. To further investigate its properties, we first construct the following model variants: (b.1) using a field-shared MLP, (b.2) removing non-linear activations, (b.3) stacking one more layer. In Fig. 5b, simplifying the *encoder* with either (b.1) or (b.2)

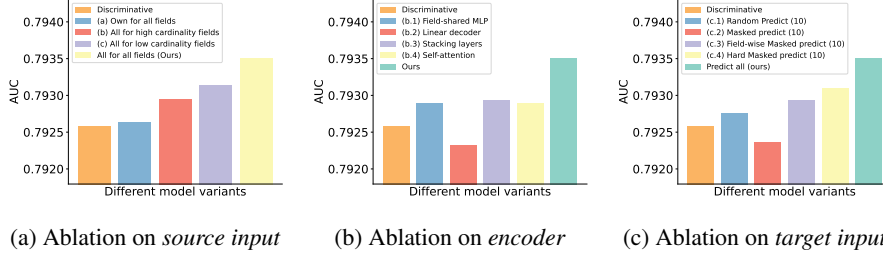


Figure 5. Ablation study on the feature generation framework design using DCN V2 on Avazu.

leads to significant performance degradation. The former underscores the importance of constructing distinct embeddings for different fields, which aligns with our intuition. The latter highlights the necessity of modeling non-linear relationships among features, which also contributes significantly to alleviating the dimensional collapse issue, as verified in Appendix C.4. This phenomenon is also observed from many other non-linear activation functions as verified in Appendix C.4. On the other hand, increasing the complexity of *encoder* with (b.3) even greatly degrades recommendation performance, AUC decreases from 0.793512 to 0.792931, which may be caused by over-fitting.

Next, we investigate whether other generative models are also feasible: (b.4) self-attention networks. In Fig. 5b, we observe that both (b.4) outperform the original discriminative paradigm but underperform the generative paradigm with the MLP-based encoder. This result confirms the effectiveness of the generative paradigm and further demonstrates that our encoder is a simple yet effective design for constructing meaningful embeddings. We conclude the following result:

Result 5. The field-wise non-linear one-layer MLP is a simple yet effective encoder. Common modifications, including simplification or increased complexity, lead to inferior recommendation performance.

Ablation on the target input design. In the proposed paradigm, we generate all feature fields simultaneously. We compare different implementations of the generative paradigm by designing distinct *target input*: (c.1) "predict-random-selected": generating only randomly selected feature fields; (c.2) "masked feature modeling": randomly masking some fields in *source input* with a learnable mask vector and predicting them as *target input*, akin to masked image modeling (He et al., 2022); (c.3) "field-aware masked feature modeling": similar to (c.2) but using field-specific mask vectors; (c.4) "hard masked feature modeling": similar to (c.2) but with zero vectors as mask. Formal definitions are detailed in Appendix B.5, and results are shown in Fig. 5c.

In the figure, all paradigms outperform the discriminative approach except (c.2), which we attribute to the superior fea-

ture distribution modeling ability of generative paradigms. For (c.2), the semantic gap between different feature fields renders the use of a single mask vector for all fields inherently impractical. Therefore, adopting (c.3) with a field-aware mask significantly improves performance. Counter-intuitively, a fixed zero vector outperforms learnable mask vectors. We hypothesize this discrepancy stems from differences between unsupervised and supervised generative paradigms. When supervised signals are introduced, the inclusion of a learnable mask vector may inadvertently impede feature distribution learning. Our "predict-all" paradigm outperforms all others, demonstrating its superiority. We conclude the following result:

Result 6. For feature generation, regarding all feature fields as target input simultaneously is effective.

4. Related Works

Feature-interaction-based recommender systems. Designing improved feature interaction models has consistently represented a significant area of research within the field of recommender systems (Zhang et al., 2019; Cheng & Xue, 2021). A key focus in the advancement of modern recommendation systems is the development of more sophisticated feature interaction modules, including first-order (Richardson et al., 2007), second-order (Rendle, 2010; Pan et al., 2018; Sun et al., 2021), and high-order interactions (Lian et al., 2018; Wang et al., 2021; Li et al., 2024). With the rise of deep learning, Deep Neural Networks (DNNs) with non-linear activation functions have been integrated into recommendation systems to capture implicit high-order feature interactions (Cheng et al., 2016; Guo et al., 2017; He & Chua, 2017; Lian et al., 2018; Wang et al., 2021). In addition to incorporating non-linearity in DNNs, several studies have explored the introduction of non-linearity in embeddings through gating mechanisms, such as FiBiNET (Huang et al., 2019), FinalMLP (Mao et al., 2023), and PEPNet (Chang et al., 2023). Orthogonal to these works, we propose a novel *Supervised Feature Generation* framework for CTR models, shifting from discriminative "feature interaction" paradigm to generative "feature generation" paradigm.

5. Conclusion

In conclusion, this work introduced a novel *Supervised Feature Generation* framework that shifts CTR modeling from discriminative feature interaction to generative feature generation. The framework’s versatility was demonstrated through reformulating various existing feature interaction models into generative ones, ranging from explicit interaction models to complex DNN-based models. It could produce feature embeddings with reduced collapse and redundancy compared to raw ID embeddings. In future work, we aim to develop a more advanced feature generation framework within the generative paradigm, incorporating enhanced encoders and other innovative components.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62441239, 62472394, U23A20319, 62441227, 62202443) as well as the Anhui Province Science and Technology Innovation Project (202423k09020011).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. The research outcomes may have various societal implications, such as optimizing online advertising strategies or enhancing e-commerce functionalities, none of which we feel must be specifically highlighted here.

References

- Avazu dataset. <https://www.kaggle.com/competitions/avazu-ctr-prediction/data>, 2014.
- Criteo dataset. <https://www.kaggle.com/c/criteo-display-ad-challenge/data>, 2014.
- Barlow, H. Redundancy reduction revisited. *Network: computation in neural systems*, 12(3):241, 2001.
- Barlow, H. B. et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961.
- Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.
- Chang, J., Zhang, C., Hui, Y., Leng, D., Niu, Y., Song, Y., and Gai, K. Pepnet: Parameter and embedding personalized network for infusing with personalized prior information. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3795–3804, 2023.
- Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10, 2016.
- Cheng, Y. and Xue, Y. Looking at ctr prediction again: Is attention all you need? In *Proceedings of the 44th international ACM SIGIR conference on research and development in information Retrieval*, pp. 1279–1287, 2021.
- Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent - A new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- Guo, H., Tang, R., Ye, Y., Li, Z., and He, X. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*, 2017.
- Guo, X., Pan, J., Wang, X., Chen, B., Jiang, J., and Long, M. On the embedding collapse when scaling up recommendation models. *ICML*, 2024.
- Harshvardhan, G., Gourisaria, M. K., Pandey, M., and Rautaray, S. S. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285, 2020.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.
- He, X. and Chua, T.-S. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval (SIGIR)*, pp. 355–364, 2017.
- Huang, T., Zhang, Z., and Zhang, J. Fibinet: combining feature importance and bilinear feature interaction for click-through rate prediction. In *Proceedings of the 13th ACM conference on recommender systems (Recsys)*, pp. 169–177, 2019.
- Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.
- Kang, W.-C. and McAuley, J. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.

- Li, H., Zhang, Y., Zhang, Y., Li, H., and Sang, L. Dcnv3: Towards next generation deep cross network for ctr prediction. *arXiv preprint arXiv:2407.13349*, 2024.
- Li, T., Chang, H., Mishra, S., Zhang, H., Katabi, D., and Krishnan, D. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023.
- Li, Z., Cui, Z., Wu, S., Zhang, X., and Wang, L. Fi-gnn: Modeling feature interactions via graph neural networks for ctr prediction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 539–548, 2019.
- Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., and Sun, G. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD)*, pp. 1754–1763, 2018.
- Mao, K., Zhu, J., Su, L., Cai, G., Li, Y., and Dong, Z. Finalmlp: An enhanced two-stream mlp model for ctr prediction. *arXiv preprint arXiv:2304.00902*, 2023.
- Oussidi, A. and Elhassouny, A. Deep generative models: Survey. In *2018 International conference on intelligent systems and computer vision (ISCV)*, pp. 1–8. IEEE, 2018.
- Pan, J., Xu, J., Ruiz, A. L., Zhao, W., Pan, S., Sun, Y., and Lu, Q. Field-weighted factorization machines for click-through rate prediction in display advertising. In *Proceedings of the 2018 world wide web conference (WWW)*, pp. 1349–1357, 2018.
- Pan, J., Xue, W., Wang, X., Yu, H., Liu, X., Quan, S., Qiu, X., Liu, D., Xiao, L., and Jiang, J. Ads recommendation in a collapsed and entangled world. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5566–5577, 2024.
- Qu, Y., Cai, H., Ren, K., Zhang, W., Yu, Y., Wen, Y., and Wang, J. Product-based neural networks for user response prediction. In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 1149–1154. IEEE, 2016.
- Rajput, S., Mehta, N., Singh, A., Keshavan, R. H., Vu, T., Heldt, L., Hong, L., Tay, Y., Tran, V. Q., Samost, J., Kula, M., Chi, E. H., and Sathiamoorthy, M. Recommender systems with generative retrieval. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- Rendle, S. Factorization machines. In *2010 IEEE international conference on data mining (ICDM)*, pp. 995–1000, 2010.
- Richardson, M., Dominowska, E., and Ragno, R. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on world wide web (WWW)*, pp. 521–530. ACM, 2007.
- Su, L., Pan, J., Wang, X., Xiao, X., Quan, S., Chen, X., and Jiang, J. Stem: Unleashing the power of embeddings for multi-task recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 9002–9010, 2024.
- Sun, H., Yu, G., Zhang, P., Zhang, B., Wang, X., and Wang, D. Graph based long-term and short-term interest model for click-through rate prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1818–1826, 2022.
- Sun, Y., Pan, J., Zhang, A., and Flores, A. Fm2: Field-matrixed factorization machines for recommender systems. In *Proceedings of the Web Conference (WWW)*, pp. 2828–2837, 2021.
- Tian, K., Jiang, Y., Yuan, Z., Peng, B., and Wang, L. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.
- Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.
- Wang, F., Gu, H., Li, D., Lu, T., Zhang, P., and Gu, N. Mcrf: Enhancing ctr prediction models via multi-channel feature refinement framework. In *International Conference on Database Systems for Advanced Applications*, pp. 359–374. Springer, 2022a.
- Wang, F., Wang, Y., Li, D., Gu, H., Lu, T., Zhang, P., and Gu, N. Enhancing ctr prediction with context-aware feature representation learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 343–352, 2022b.
- Wang, F., Gu, H., Li, D., Lu, T., Zhang, P., Shang, L., and Gu, N. A comprehensive summarization and evaluation of feature refinement modules for ctr prediction. *arXiv preprint arXiv:2311.04625*, 2023.

- Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., and Chi, E. Dcn-v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference (WWW)*, pp. 1785–1797, 2021.
- Wang, Y., Qin, Y., Sun, F., Zhang, B., Hou, X., Hu, K., Cheng, J., Lei, J., and Zhang, M. Disenctr: Dynamic graph-based disentangled representation for click-through rate prediction. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 2314–2318, 2022c.
- Yin, M., Wang, H., Guo, W., Liu, Y., Zhang, S., Zhao, S., Lian, D., and Chen, E. Dataset regeneration for sequential recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3954–3965, 2024.
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021.
- Zhai, J., Liao, L., Liu, X., Wang, Y., Li, R., Cao, X., Gao, L., Gong, Z., Gu, F., He, J., Lu, Y., and Shi, Y. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- Zhang, S., Yao, L., Sun, A., and Tay, Y. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.
- Zhang, W., Du, T., and Wang, J. Deep learning over multi-field categorical data: A case study on user response prediction. In *Advances in information retrieval: 38th European conference on IR research (ECIR)*, pp. 45–57, 2016.
- Zhu, J., Liu, J., Yang, S., Zhang, Q., and He, X. Fuxictr: An open benchmark for click-through rate prediction. *arXiv preprint arXiv:2009.05794*, 2020.
- Zhu, J., Dai, Q., Su, L., Ma, R., Liu, J., Cai, G., Xiao, X., and Zhang, R. Bars: Towards open benchmarking for recommender systems. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval (SIGIR)*, pp. 2912–2923, 2022.

A. Formal reformulation of existing feature interaction models

In Tab. 2, we provide the formal definition of how to reformulate existing discriminative models into generative paradigms. Notably, we only present the "feature interaction" or "feature generation" part in each paradigm for simplicity.

Table 2. Feature interaction models in discriminative & generative paradigm

Model	Discriminative	Generative
FM	$\sum_{i,j=1}^N \mathbf{v}_j \odot \mathbf{v}_i$	$\sum_{i,j=1}^N \sigma([\mathbf{v}] \cdot W_{F(j)}) \odot \mathbf{v}_i$
FmFM	$\sum_{i,j=1}^N \mathbf{v}_j \odot [\mathbf{v}_i \cdot M_{F(i) \rightarrow F(j)}]$	$\sum_{i,j=1}^N \sigma([\mathbf{v}] \cdot W_{F(j)}) \odot [\mathbf{v}_i \cdot M_{F(i) \rightarrow F(j)}]$
CrossNet V2	$\sum_{l=1}^L \sum_{i,j=1}^N \mathbf{v}_j^0 \odot (\mathbf{v}_i^l \cdot M_{F(i) \rightarrow F(j)}^l)$	$\sum_{l=1}^L \sum_{i,j=1}^N \sigma([\mathbf{v}]^l \cdot W_{F(j)}^l) \odot (\mathbf{v}_i^l \cdot M_{F(i) \rightarrow F(j)}^l)$
DeepFM	$\sum_{i,j=1}^N \mathbf{v}_j \odot \mathbf{v}_i + \text{DNN}([\mathbf{v}])$	$\sum_{i,j=1}^N \sigma([\mathbf{v}] \cdot W_{F(j)}) \odot \mathbf{v}_i$
xDeepFM	$\sum_{l=1}^L \sum_{i,j=1}^N \text{Conv}^l(\mathbf{v}_j^0 \odot \mathbf{v}_i^l) + \text{DNN}([\mathbf{v}])$	$\sum_{l=1}^L \sum_{i,j=1}^N \text{Conv}^l(\sigma([\mathbf{v}]^l \cdot W_{F(j)}^l) \odot \mathbf{v}_i^l) + \text{DNN}([\mathbf{v}])$
IPNN	$\text{DNN}([\mathbf{v}], \sum_{i,j=1}^N \mathbf{v}_j \odot \mathbf{v}_i)$	$\text{DNN}([\mathbf{v}], \sum_{i,j=1}^N \sigma([\mathbf{v}] \cdot W_{F(j)}) \odot \mathbf{v}_i)$
DCN V2	$\sum_{l=1}^L \sum_{i,j=1}^N \mathbf{v}_j^0 \odot (\mathbf{v}_i^l \cdot M_{F(i) \rightarrow F(j)}^l) + \text{DNN}([\mathbf{v}])$	$\sum_{l=1}^L \sum_{i,j=1}^N \sigma([\mathbf{v}]^l \cdot W_{F(j)}^l) \odot (\mathbf{v}_i^l \cdot M_{F(i) \rightarrow F(j)}^l) + \text{DNN}([\mathbf{v}])$

B. Detailed experimental configuration

B.1. Dataset statistics

We adopt the Criteo x1 and Avazu x4 datasets provided by FuxiCTR (Zhu et al., 2020; 2022), whose statistics are summarized in Tab. 3.

B.2. Implementation details of baseline methods

We first introduce common settings for all models: (1) For Criteo dataset, the embedding size is set to 10, batch size is set to 4,096, and learning rate is set to 1e-3. (2) For Avazu dataset, the embedding size is set to 16, batch size is set to 10,000, and learning rate is set to 1e-3. All experiments will be early stopped when results on validation dataset decrease for consecutive two training epochs.

Then we list the detailed setting of different baseline models. Notably, we do not tune these hyper-parameters when fitting these models into the proposed generative paradigm:

- FM: embedding regularization coefficient is set to 5.0e-06 for Criteo and 1.0e-06 for Avazu.
- FmFM: parameter regularization coefficients are set to 1.0e-06 for the both datasets; we adopt matrixed field embedding transform type (Sun et al., 2021) for both datasets.
- CrossNet V2: embedding regularization coefficient is set to 1.0e-05 and 0 for Criteo and Avazu, respectively; number of cross layers is set to 3, 5 for Criteo and Avazu.
- DeepFM: embedding regularization coefficient is set to 1.0e-05 and 0 for Criteo and Avazu, respectively; a parallel DNN with size [400, 400, 400] and [2000, 2000, 2000, 2000] are used for Criteo and Avazu, respectively.

Table 3. The number of user-item interactions of the adopted two datasets.

	Train	Valid	Test
Criteo	33M	8M	4M
Avazu	32M	4M	4M

- xDeepFM: embedding regularization coefficient is set to 1.0e-05 and 0 for Criteo and Avazu, respectively; CIN hidden units are set to [16, 16] and [276] for Criteo and Avazu, respectively; DNN size is set to [400, 400, 400] and [500, 500, 500] for Criteo and Avazu, respectively.
- IPNN: embedding regularization coefficient is set to 1.0e-05 and 1.0e-09 for Criteo and Avazu, respectively; DNN size is set to [400, 400, 400] and [1000, 1000, 1000] for Criteo and Avazu, respectively.
- DCN V2: based on the setting of CrossNet V2, a parallel DNN with size [500, 500, 500] and [2000, 2000, 2000, 2000] are used for Criteo and Avazu, respectively.

All experiments can fit into a GPU with 14GB memories.

B.3. Computational complexity analysis

Table 4. Computational complexity when reformulating a discriminative feature interaction model into a generative feature generation model. The proposed generative paradigm achieves significant recommendation performance improvements, as detailed in Section 3.2, while incurring only a marginal increase in computational overhead—averaging 3.14% more computation time and 1.45% additional GPU memory consumption.

Model		Criteo		Avazu	
		Speed (time/epoch)	GPU memory (MB)	Speed (time/epoch)	GPU memory (MB)
Explicit	FM	DIS	8m40s	3m08s	2554
		GEN	8m58s	3m13s	2626
	FmFM	DIS	9m38s	3m48s	10148
		GEN	9m45s	3m57s	10180
	CrossNetv2	DIS	5m01s	1m43s	2100
		GEN	5m09s	1m57s	2190
DNN-based	DeepFM	DIS	8m23s	4m36s	3622
		GEN	8m33s	4m44s	3676
	xDeepFM	DIS	8m15s	2m43s	3268
		GEN	8m24s	2m45s	3322
	IPNN	DIS	6m37s	3m06s	2592
		GEN	6m41s	3m14s	2614
	DCNv2	DIS	5m31s	5m44s	2982
		GEN	5m58s	6m01s	3070

Assuming the original model in the discriminative paradigm has complexity $O(A)$, the primary computational overhead when transitioning to a generative paradigm arises from the *encoder*. The encoder is implemented as a field-wise non-linear MLP, formally defined as:

$$f_{\text{encoder}}^i([v]) = \sigma([v]W_{F(i)}), \quad (7)$$

where $[v] \in \mathbb{R}^{NK}$ denotes the concatenation of all feature embeddings, and $W_{F(i)} \in \mathbb{R}^{NK \times K}$ represents a field-wise weight matrix. Consequently, the total encoder complexity becomes $O(BLN^2d^2)$, where B is the batch size, L denotes the number of encoder layers, N the number of feature fields, and d the embedding dimension. This computational complexity aligns with mainstream discriminative feature interaction models (e.g., DCN V2), indicating comparable efficiency. Furthermore, as demonstrated in the *source input* ablation study (Sec. 3.4), the complexity can be reduced to $O(BLN'^2d^2)$ by using all field embeddings as *source input* only for fields with the lowest cardinality, achieving this optimization with moderate performance trade-offs.

Result 7. The extra computational burden introduced by reformulating existing discriminative feature interaction paradigms to the generative feature generation paradigm is marginal.

B.4. Robustness analysis of the batch-wise setting in Sec. 3.3.1.

In Sec. 3.3.1, we have conducted embedding analyses in dimensional collapse with a batch-wise setting, which greatly accelerates the analysis process compared with that based on the full validation dataset. But this batch-wise setting may introduce randomness to the analysis results, so we further provided the analysis of different seeds in Fig. 6. In the figure, the trend of embedding spectra is consistent across all seeds, demonstrating the robustness of our batch-wise analysis setting. Specifically, on both Avazu and Criteo, the spectrum curves of discriminative paradigms exhibit an abrupt singular decay from 1×10^{-5} to 1×10^{-15} , a reduction of 10^{10} times. This indicates a severe dimensional collapse issue. But in our generative paradigm, the abrupt singular value decay has been greatly alleviated. This verifies that the generative paradigm substantially mitigates the embedding dimensional collapse issue, forming a more balanced embedding space.

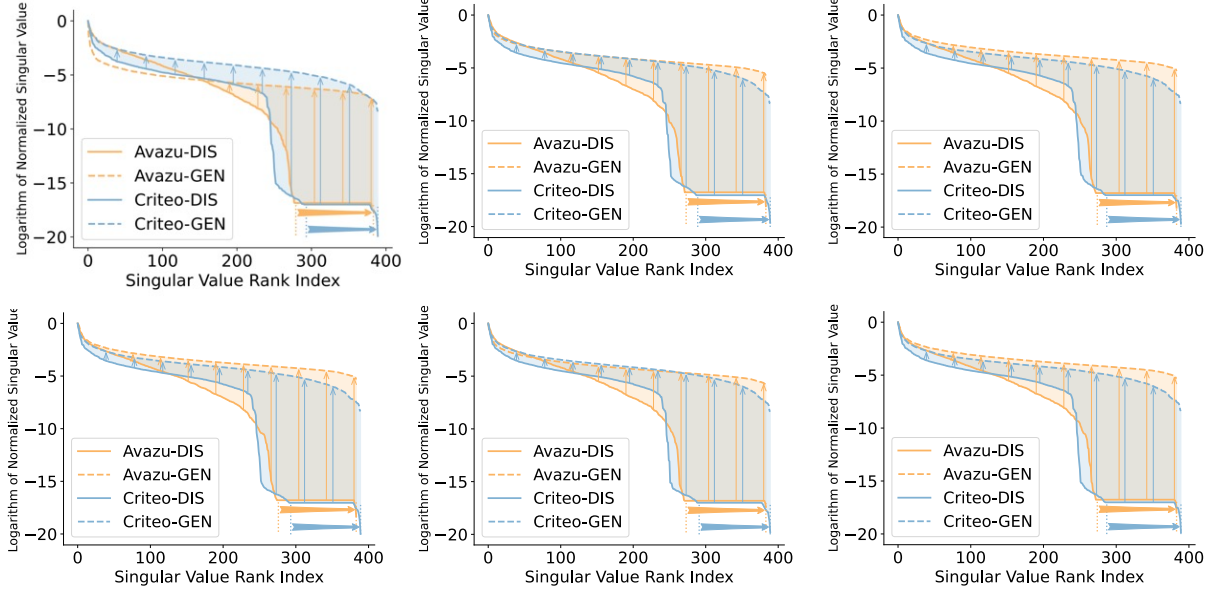


Figure 6. Normalized embedding spectrum visualization with batch-wise setting in different seeds. We can observe that the trend of embedding spectra is consistent across all seeds, which demonstrates the robustness of our batch-wise analysis setting.

B.5. Formal definition of different *target input* design

We provide a detailed formal definition of the different *target input* designs mentioned in Sec. 3.4.

(c.1) "Predict-random-selected", which generates only randomly selected feature fields:

$$y = \sum_i \sum_{j \in \mathcal{F}_{random}} \sigma([v] \cdot W_{F(i)}) \odot v_j, \quad (8)$$

where $\mathcal{F}_{random}$ is a set of fields randomly sampled from all fields.

(c.2) "Masked feature modeling": randomly masking some fields in *source input* with a learnable mask vector and predicting them as *target input*, akin to masked image modeling (He et al., 2022):

$$y = \sum_i \sum_{j \in (\mathcal{F}_{unmask} \cup \mathcal{F}_{mask})} \sigma([v]_{\text{not masked}} \cdot W_{F(i)}) \odot v_{j,mask} \quad (9)$$

where $\mathcal{F}_{unmask} \cup \mathcal{F}_{mask} = \mathcal{F}$, $v_{j,mask} = \text{mask}$ if j in $\mathcal{F}_{mask}$ else v_j .

(c.3) "Field-aware masked feature modeling": similar to (c.2) but using field-specific mask vectors:

$$y = \sum_i \sum_{j \in (\mathcal{F}_{unmask} \cup \mathcal{F}_{mask})} \sigma([v]_{\text{not masked}} \cdot W_{F(i)}) \odot v_{j,mask} \quad (10)$$

where $\mathcal{F}_{unmask} \cup \mathcal{F}_{mask} = \mathcal{F}$, $v_{j,mask} = \text{mask}_j$ if j in $\mathcal{F}_{mask}$ else v_j .

(c.4) "Hard masked feature modeling": similar to (c.2) but with zero vectors as mask:

$$y = \sum_i \sum_{j \in (\mathcal{F}_{unmask} \cup \mathcal{F}_{mask})} \sigma([v]_{\text{not masked}} \cdot W_{F(i)}) \odot v_{j,mask} \quad (11)$$

where $\mathcal{F}_{unmask} \cup \mathcal{F}_{mask} = \mathcal{F}$, $v_{j,mask} = \mathbf{0}$ if j in $\mathcal{F}_{mask}$ else v_j .

C. Supplemental results

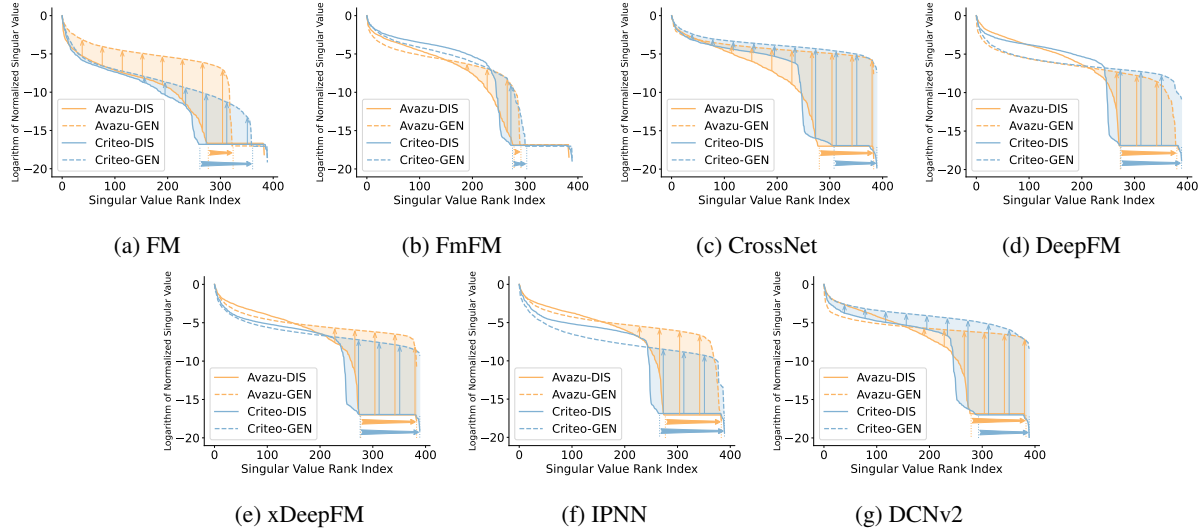


Figure 7. Normalized singular value spectrum of embeddings used to interact with raw ID embeddings. It is the concatenation of raw ID embeddings for the discriminative paradigm, while the embedding immediately constructed by the *encoder* for the generative paradigm.

C.1. Normalized singular value spectrum visualization of all models

The normalized singular value spectrum of all models are illustrated in Fig. 7. Similar to results concluded in Sec. 3.3.1, the feature generation framework substantially mitigates the embedding dimensional collapse issue, forming a more balanced and meaningful embedding space.

C.2. Dimensional collapse analysis of embedding lookup tables

In Sec. 3.3.1, we focus on analyzing the spectrum of embeddings used to interact with the original embeddings, since we are mainly motivated to address the dimensional collapse issue of these embeddings. On the other hand, we can also follow Guo et al. (2024) to visualize the spectrum of embedding lookup tables, *i.e.*, $V_i \in \mathbb{R}^{D_i \times K}$ defined in Sec. 2.1, where i denotes one of the feature field, D_i is the field’s cardinality, and K is the embedding dimension size of the embedding table. The results have been depicted in Fig. 8. In the figure, the spectrum of high-cardinality embedding lookup tables in the generative paradigm is higher than the discriminative one. This indicates the embedding space will be less dominated by some specific dimensions, which will greatly enhance the robustness of these embeddings. However, for those low-cardinality embeddings, the improvement remains limited. This is fundamentally because these field embeddings are inherently constrained by nature. For instance, the number of meaningful singular values of a matrix sized $4 \times K$ cannot exceed four.

C.3. Pearson correlation matrix of all models

Similar to Sec. 3.3.2, we provide Pearson correlation matrix of all models on the Avazu dataset in Fig. 12. The conclusion remains the same as in Sec. 3.3.2: (1) There is a strong connection between redundancy reduction metric and recommendation performance: The most simple model FM yields the most matrix with intra-field and inter-field correlations, while the correlation matrix of other models are reduced to some extent, depending on whether DNN (DeepFM, IPNN) or more

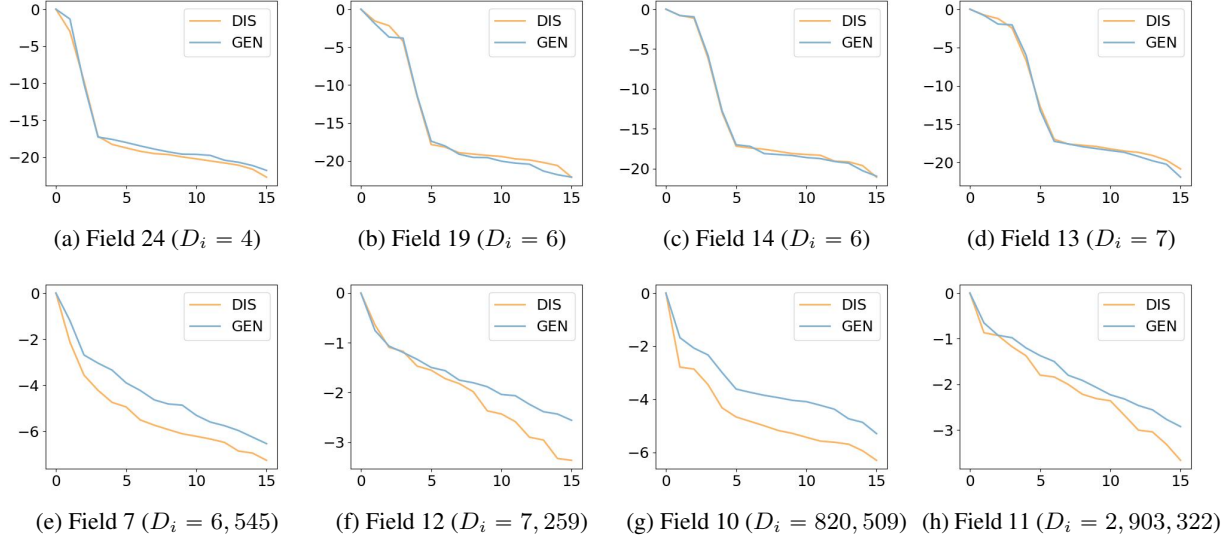


Figure 8. Normalized singular value spectrum of embeddings lookup tables $\mathbf{V}_i \in \mathbb{R}^{D_i \times K}$, where i denotes one of the feature field, D_i is the field’s cardinality, and K is the embedding dimension size of the embedding table.

advanced interaction modules (CrossNet V2, xDeepFM, DCN V2). (2)The feature generation framework produces embeddings highly de-correlated with raw ID embeddings: We can observe that the correlation matrices of all models become a nearly zero matrix within the generative paradigm.

C.4. Comparison of different non-linear activation functions

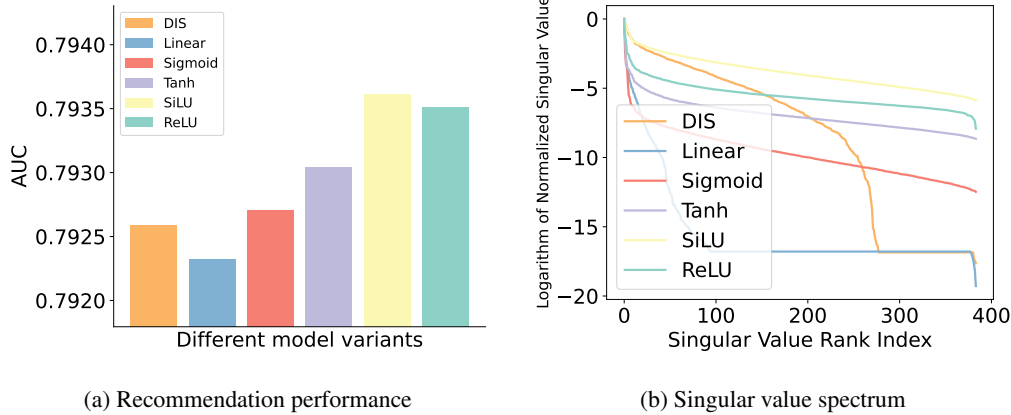


Figure 9. We have implemented the *encoder* with different non-linear activation functions, including *ReLU*, *Sigmoid*, *Tanh*, and *SiLU*, and providing the corresponding results based on DCN V2: (a) The recommendation performance with different non-linear activation functions. (b) The normalized singular value spectrum of the embedding space with different non-linear activation functions.

We employ a field-wise non-linear single-layer MLP as our *encoder*, with the non-linear activation function being one of its most critical components. A natural question arises regarding the role of the non-linear activation function and the criteria for selecting an appropriate one. We have empirically assessed the effects of various activation functions on the *encoder*, with the findings illustrated in Fig. 9. As depicted in Fig. 9a, the absence of a non-linear activation function in the *encoder* results in a notable decline in performance, underscoring the importance of incorporating non-linearity within the *encoder*. Conversely, all non-linear activation functions enhance recommendation performance relative to the discriminative paradigm, with the rank of recommendation performance being *Sigmoid* < *Tanh* < *ReLU* < *SiLU*. Furthermore, we present the normalized singular value spectrum of embeddings in Fig. 9b. Initially, the spectrum of the linear activation is highly

collapsed, potentially accounting for its inferior recommendation performance. Subsequently, it is observable that the spectra of all non-linear activation functions exhibit greater smoothness than that of the discriminative one. This suggests that non-linear activation functions play a pivotal role in alleviating the embedding dimensional collapse issue. Additionally, the spectrum adheres to the rank $Sigmoid < Tanh < ReLU < SiLU$, mirroring the ranking of recommendation performance. This observation further implies a strong correlation between the mitigation of embedding dimensional collapse and the enhancement of recommendation performance.

Result 8. The non-linear activation function is an important component of the field-wise MLP encoder, crucial for embedding dimensional collapse mitigation. Besides, many non-linear activation functions, including Sigmoid, ReLU, Tanh, and SiLU, can get consistent performance lift while mitigating the dimensional collapse.

C.5. Comparison with feature refinement and graph-based models

Some other methods also target enhancing the embeddings of CTR models with field graphs (Sun et al., 2022; Li et al., 2019; Wang et al., 2022c) or feature enhancement modules (Wang et al., 2023; 2022b;a). Our paradigm differs from these works in the sense that we aim to tackle the dimensional collapse issue due to the direct interaction of ID embeddings. We have empirically compared our paradigm with several representative feature refinement models, with results depicted in Tab. 5. We observed that some models outperform the discriminative DCNv2 models, but still underperform our generative paradigm. Besides, we also studied the singular spectrum in Fig. 10, and we find that the feature enhancement methods can mitigate the dimensional collapse on the tail singular values compared to the vanilla discriminative DCN V2. However, our generative paradigm leads to more robust values across all dimensions.

Table 5. Comparison with other methods that also target enhancing embeddings of CTR models. We have compared with one classic field-graph method Fi-GNN (Li et al., 2019), and two feature enhancement methods GFRL (Wang et al., 2022a) and FRNet (Wang et al., 2022b). Notably, we also visualize the normalized spectrum of these methods in Fig.

Model		Criteo		Avazu	
		AUC↑	Logloss↓	AUC↑	Logloss↓
FiGNN		0.81352	0.43845	0.79156	0.37343
DCNv2	DIS	0.81387	0.43826	0.79282	0.37222
	GFRL	0.81427	0.043773	0.79296	0.37194
	FRNet	0.81431	0.43789	0.79313	0.37191
	GEN	0.81472	0.43713	0.79342	0.37180

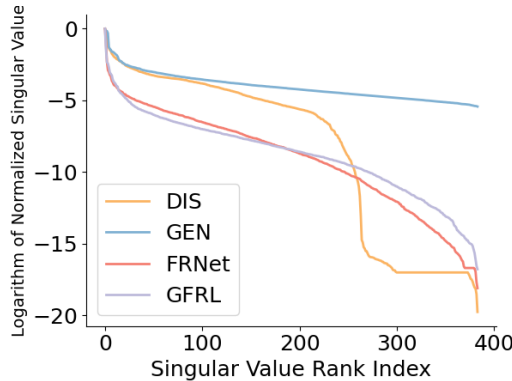


Figure 10. Normalized embedding spectrum of the feature enhancement methods. We can find that these feature enhancement methods can mitigate the dimensional collapse on the tail singular values compared to the vanilla discriminative DCN V2. However, our generative model leads to more robust values on all dimensions.

C.6. T-SNE visualization comparison

In Fig. 11, we have visualized discriminative and generative embeddings with different cardinalities with T-SNE (Van der Maaten & Hinton, 2008). Fig. 11d and Fig. 11h depict embeddings of the highest cardinality field in the dataset, where we

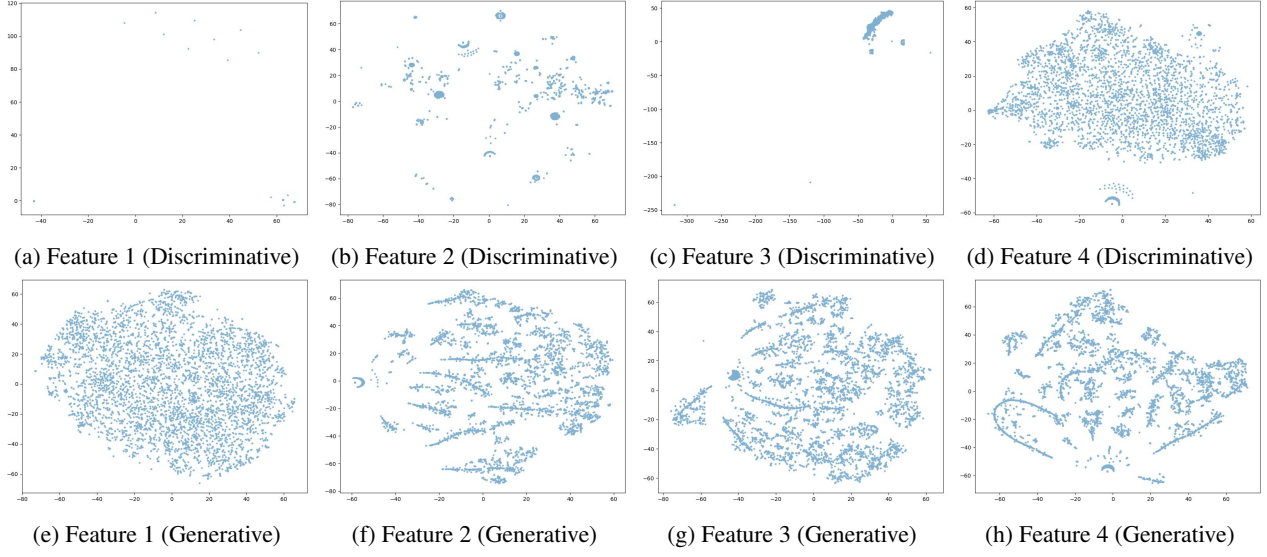


Figure 11. T-SNE visualisation of discriminative and generative embeddings of four features, numbered from 1 to 4. The cardinality of these features is 4, 4,051, 820,509, and 2,903,322, respectively. (a-d) illustrate embeddings of the four features within the discriminative paradigm; (e-h) illustrate embeddings of the four features within the generative paradigm.

observe that the generative embeddings retain the separability as the discriminative paradigm. However, the improvement brought by the generative paradigm is substantial for embeddings of fields with less cardinality. In Fig. 11a, Fig. 11b, and Fig. 11c, the embeddings coalesce in the latent space, even for the field with the second-highest cardinality (Fig. 11c and Fig. 11g). After the generative reformulation, all three embeddings can form a more uniform distribution in the latent space, as illustrated respectively in Fig. 11e, Fig. 11f, and Fig. 11g. These results demonstrate that our generative paradigm can greatly improve the separability of embeddings, especially for embeddings with fewer cardinalities. This also supplements the aforementioned dimensional collapse phenomena analysis from a field-wise perspective.

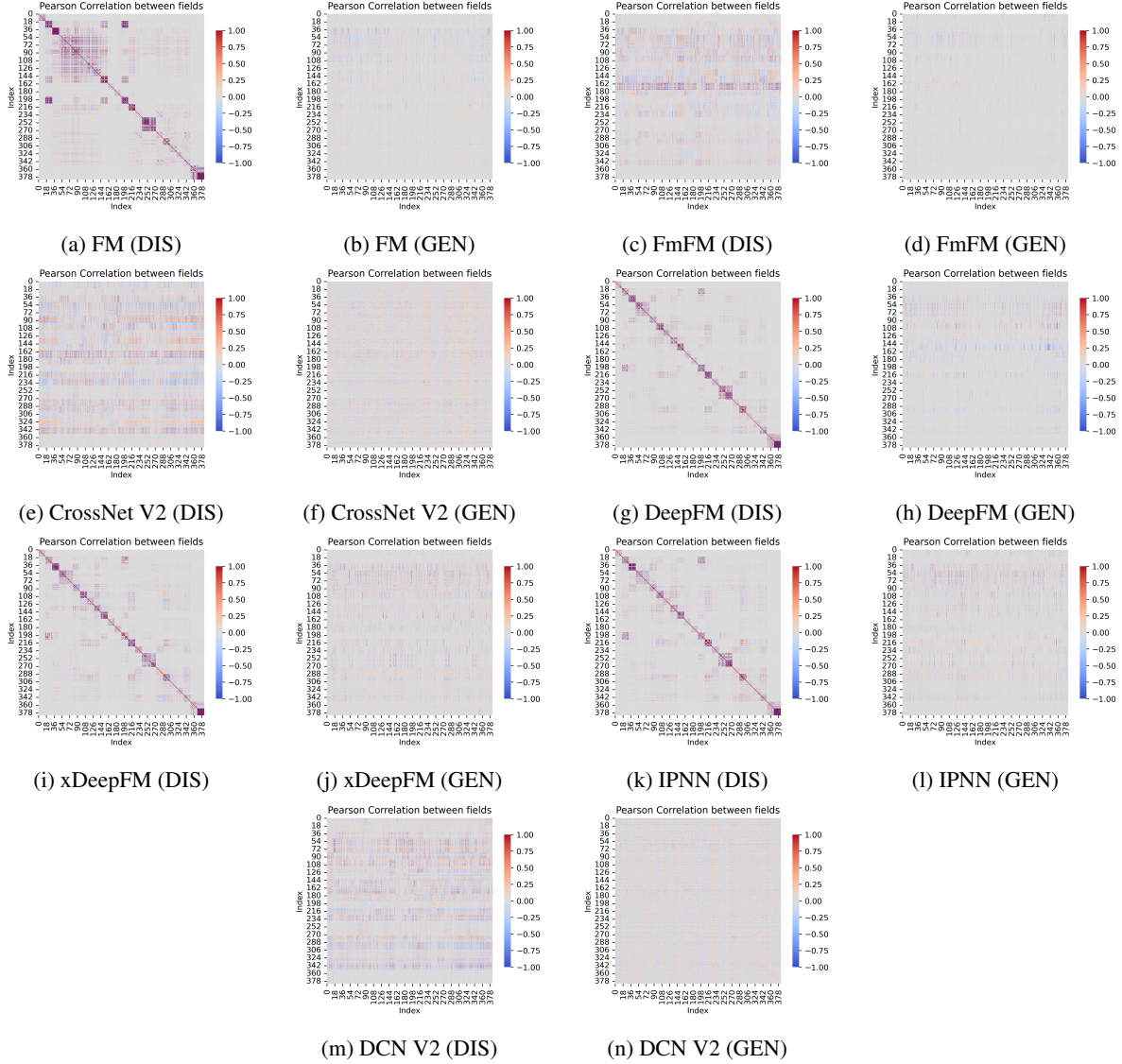


Figure 12. Pearson correlation matrix between two interacted embeddings. For all discriminative feature interaction models, the correlation matrix becomes a nearly zero matrix after reformulating them into a generative paradigm, which perfectly aligns with the redundancy reduction principle.