

Beyond Bag-of-Words: Generative Structural Regularization for Dense Table Retrieval

Weijie Liu¹, Shanshan Chen¹, Fang Kong^{1,2*}

¹School of Computer Science and Technology, Soochow University, China

²Jiangsu Key Lab of Language Computing, Suzhou 215123, China

weijieliu.cs@gmail.com, kongfang@suda.edu.cn

Abstract

Dense retrieval has become the dominant paradigm for large-scale table retrieval. However, existing contrastive retrievers often suffer from bag-of-words degradation: they rely on lexical overlap rather than row-column bindings. To quantify and mitigate this, we propose an adversarial benchmark and a structure-aware retrieval framework. Our benchmark uses “evidence cell” anchoring and a *Homologous Structural Swapping* mechanism to construct lexically equivalent but structurally invalid negative samples. Our model uses text generation as a structural regularizer for dense table representations. It decouples row and column anchors through orthogonal attention pooling, and employs dynamic masking to encourage autoregressive reconstruction. This process encourages the retrieval representations to retain table content and structure-related cues. Extensive experiments on OTT-QA, NQ-Tables, and our adversarial benchmark demonstrate the effectiveness of our method. It achieves competitive performance on conventional retrieval datasets and substantially improves structural robustness under adversarial settings.

1 Introduction

Tables encode structured knowledge and serve as key resources for NLP tasks such as open-domain question answering (Chen et al., 2020) and fact verification (Aly et al., 2021). Recently, dense retrieval models have become the mainstream paradigm for locating tables within large-scale corpora (Singh et al., 2025). As illustrated on the left side of Figure 1, existing methods typically employ a standard dual-encoder architecture (Herzig et al., 2021; Wang et al., 2022). They map queries and tables into a unified vector space via contrastive learning, enabling efficient end-to-end retrieval.

However, the existing contrastive learning paradigm faces critical challenges when applied to tabu-

lar data. Under objectives such as InfoNCE (Oord et al., 2018), models may exploit shortcut learning (Geirhos et al., 2020) rather than capture table structure. This issue is amplified by prevailing table retrieval benchmarks, such as OTT-QA (Chen et al., 2020) and NQ-Tables (Herzig et al., 2021), where negative samples often differ lexically from the ground-truth tables. As a result, models can minimize empirical risk by relying on local word overlap instead of modeling row-column dependencies. As illustrated on the right side of Figure 1, a table with substantial lexical overlap but scrambled logical relations can still receive a high similarity score. This reveals a severe bag-of-words degradation (Jia and Liang, 2017): models appear to achieve high recall while remaining insufficiently sensitive to row-column bindings.

To quantify this degradation phenomenon and mitigate the over-reliance on lexical features, we propose solutions from two dimensions: the evaluation benchmark and the representation learning framework. First, we construct a novel adversarial table retrieval benchmark with explicit row-column binding constraints. By introducing “evidence cell” anchoring and a *Homologous Structural Swapping* mechanism, we generate adversarial negative samples that exhibit lexical equivalence but inconsistent row-column bindings. Under this benchmark, retrievers that rely primarily on bag-of-words matching suffer substantial performance degradation.

Second, to improve sensitivity to row-column bindings, we propose Generative Structural Regularization for Table retrieval (**GSR-Table**). Our core insight is to use text generation as a structural regularizer for dense table representations. Specifically, we introduce orthogonal attention pooling to decouple row- and column-level anchors, denoted as z_{row} and z_{col} . We further design a generative information bottleneck via *Dynamic Masking*. This encourages the decoder to perform autoregressive

*Corresponding author.

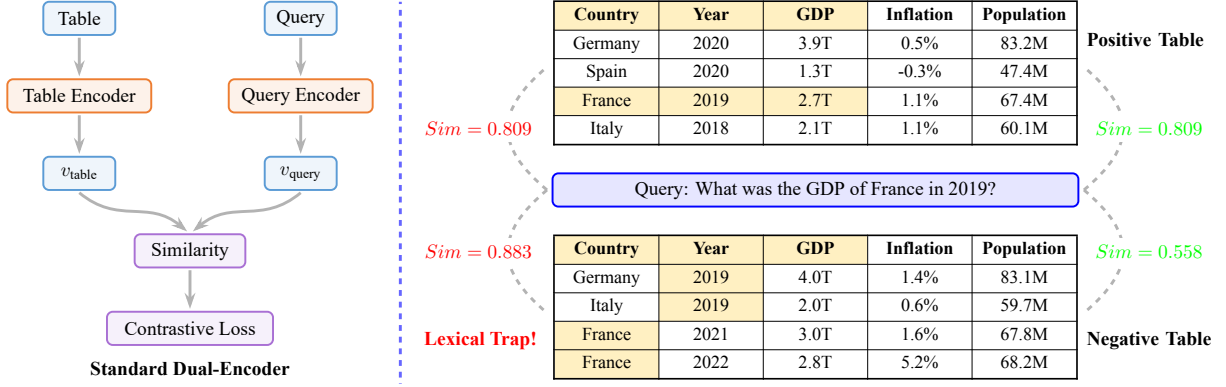


Figure 1: **Left:** The standard dual-encoder architecture based on contrastive learning. **Right:** Illustration of the bag-of-words degradation phenomenon in table retrieval. Existing methods (red scores) are misled by lexically overlapping negative tables, whereas our GSR-Table (green scores) better resists this lexical trap.

reconstruction from masked representations, placing reconstruction pressure on the multi-view representations through a *Structural Compensation* mechanism. Ultimately, this process encourages the retrieval vectors to retain table content and cues about row-column associations.

We conduct extensive experiments on the OTT-QA and NQ-Tables datasets, as well as our newly constructed adversarial benchmark. Results demonstrate that GSR-Table achieves competitive performance on conventional retrieval datasets. It also exhibits superior robustness on the adversarial benchmark, indicating greater sensitivity to perturbations of row-column bindings. We release the code and data for GSR-Table.¹

Our contributions can be summarized as follows: (1) We identify bag-of-words degradation in table retrieval, where contrastive retrievers rely on lexical overlap rather than table structure. (2) We propose Homologous Structural Swapping to construct lexically equivalent but structurally invalid negatives for adversarial evaluation. (3) We introduce GSR-Table, which uses generative reconstruction with dynamic masking and structural compensation to improve structure-aware dense representations.

2 Related Work

2.1 Table Retrieval

Table retrieval aims to retrieve query-relevant tables from large-scale corpora. Early methods relied on surface-level signals, such as lexical matching, heuristic ranking features, or entity linking (Zhang and Balog, 2018; Cafarella et al., 2008; Bhagavatula et al., 2015). With advances in pre-trained

language models, dense retrieval based on dual encoders and contrastive learning has become a dominant paradigm (Herzig et al., 2021; Wang et al., 2022). However, compressing a table into a single dense vector makes it difficult to preserve its complex two-dimensional structure. To alleviate this, THYME (Li et al., 2025) combines dense semantic and sparse lexical matching, and uses a field-aware mechanism to capture field-specific preferences. SSDR (Jin et al., 2023) constructs phrase-level query representations and header/value-level table representations, enabling more fine-grained matching.

Although THYME and SSDR improve the field modeling and fine-grained matching, they largely overlook the optimization biases induced by contrastive learning. Following this line, we also adopt a dual-encoder framework but focus on whether retrievers truly use table structure rather than lexical shortcuts. To this end, we analyze bag-of-words degradation in table retrieval and introduce generative structural regularization to encourage retrieval vectors to retain structural information.

2.2 Structure-Aware Table Representations

To model the two-dimensional structure of tables, prior studies have proposed various structure-aware table representation methods. Representative methods inject encoder-side structural signals, such as row and column embeddings (Herzig et al., 2020), structured attention (Liu et al., 2025), header-cell relations (Kale and Rastogi, 2020), or table-specific pre-training objectives (Alonso et al., 2024). These methods show that explicitly encoding table layouts and cell dependencies can improve performance on table understanding tasks. However, in table retrieval, the final decision relies on low-dimensional

¹<https://github.com/weijieliu-cs/GSR-Table>

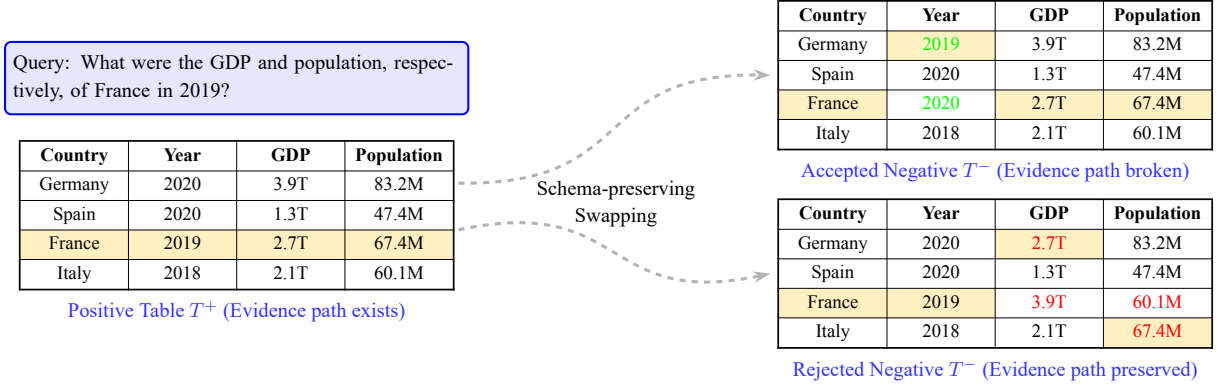


Figure 2: Illustration of structure-adversarial negative construction through schema-preserving swapping. The positive table T^+ supports the evidence path France $\rightarrow$ 2019 $\rightarrow$ {GDP, Population}. Accepted negatives break this path and become unanswerable, while rejected negatives still preserve it and remain relevant.

dense vectors obtained through pooling. Without direct constraints on the final retrieval vectors, structural information captured by the encoder may be weakened during contrastive optimization.

Unlike prior encoder-side methods, GSR-Table introduces decoder-side structural regularization. Specifically, orthogonal attention pooling decouples row- and column-level anchors, while dynamic masking creates a generative bottleneck for autoregressive reconstruction. This bottleneck places reconstruction pressure on the multi-view representations, thereby encouraging the final dense vectors to retain row-column binding cues.

3 The Adversarial Structural Table Retrieval Benchmark

To assess whether retrieval models are sensitive to row-column bindings beyond lexical overlap, we construct a structure-adversarial benchmark for table retrieval. Unlike datasets such as OTT-QA and NQ-Tables, our benchmark enforces *Bag-of-Words Equivalence* and *Structural Dissonance*, requiring structure-aware retrieval beyond lexical matching.

3.1 Query Generation with Fine-grained Evidence Anchoring

Specifically, we clean and randomly sample 68K high-quality tables from the WDC corpus (Lehmberg et al., 2016). For each table, we use Gemini-3.1-flash (Google DeepMind, 2026a) to generate a query-answer pair, and require the model to perform *Fine-grained Evidence Anchoring*. For each query q , the model outputs both the answer a and the supporting evidence cells $\mathcal{E}_q = \{(i_k, j_k)\}_{k=1}^K$, where (i_k, j_k) denotes the two-dimensional physical coordinate of an evidence cell in table T . This

design provides each query with an explicit structural evidence path, which serves as an operation anchor for subsequent adversarial negative construction. To reduce generation bias and hallucination noise, we further use GPT-5-mini (OpenAI, 2025) to verify the consistency of each generated $(q, \mathcal{E}_q, a)$ triple and filter out invalid samples. This verification rejects 4,791 triples and retains 63,209 valid examples.

3.2 Schema-preserving Homologous Structural Swapping

For each positive test pair (T^+, q) , we exploit the schema-preserving property of tables to synthesize 10 adversarial negatives $\{T_1^-, \dots, T_{10}^-\}$. Specifically, in relational tables, cells in the same column are typically governed by the same attribute definition, while cells in the same row usually belong to the same entity or record. Therefore, tabular facts are not merely unordered collections of cell contents, but are determined by the orthogonal dependencies induced by two-dimensional row-column coordinates.

Based on this property, we propose *Homologous Structural Swapping*. Given an evidence cell $c_{i,j} \in \mathcal{E}_q$, if the first row serves as headers, all cells in column j share the same attribute semantics. We then uniformly sample another non-evidence cell $c_{i',j}$ from the same column, where $c_{i,j} \neq c_{i',j}$, and swap their textual contents to obtain a perturbed table T^- . Since the swap occurs within the same attribute column, it preserves the table schema while breaking the evidence cell’s binding to its original row entity. If the first column serves as headers, we perform the symmetric operation along the row dimension. For queries involving multiple evidence

cells, we allow joint swapping over any subset of $\mathcal{E}_q$ to further strengthen the perturbation. The construction process is illustrated in Figure 2.

This construction gives adversarial negatives a mathematical property. For any positive table T^+ and its corresponding negative table T^- , we have:

$$\text{BoW}(T^+) = \text{BoW}(T^-). \quad (1)$$

That is, the positive and negative tables contain the same lexical items and values, eliminating OOV signals and surface-level lexical differences. Therefore, retrieval models relying on keyword matching or shallow lexical co-occurrence cannot distinguish them. Only models capturing row-entity, column-attribute, and cell-content bindings can identify the structural-semantic dissonance in T^- .

3.3 Closed-loop Answerability Verification

To reduce false negatives introduced by homologous swapping, we introduce a closed-loop answerability verification procedure. In table retrieval, the core objective is to assess the relevance of a table to a query, rather than the correctness of a specific answer. For example, a swap may only change the value of the target attribute. In this case, the perturbed table T^- may still support the original query q , and thus remains relevant under the retrieval setting. Accordingly, a candidate T^- is valid only if the swap breaks the two-dimensional mapping logic required by q , making T^- unanswerable.

To verify this, we feed each candidate T^- into Gemini-3.1-flash for reverse inference. We instantiate the LLM as a Boolean evaluator under a zero-shot setting, requiring it to determine whether q can be answered from the existing structure of T^- . Formally, we define the answerability function $\mathcal{M}_{\text{eval}}$ as:

$$\mathcal{M}_{\text{eval}}(q, T^-) \in \{0, 1\}, \quad (2)$$

where 1 indicates that the table contains a complete structural logic supporting the query, while 0 indicates that structural dissonance makes the query unanswerable. A candidate adversarial negative is accepted only when:

$$\mathcal{M}_{\text{eval}}(q, T^-) = 0. \quad (3)$$

This closed-loop verification reduces false negatives that merely alter factual values while preserving valid query-supporting structures. As a result, the benchmark focuses on evaluating whether retrieval models can identify valid row-column bindings rather than exploit shallow lexical shortcuts.

We use a test-first split requiring ten verified HSS negatives per example. Further details and statistics are provided in Appendix A.1.

Finally, to simulate the realistic open-domain retrieval, we further introduce 320K query-irrelevant tables from the WDC corpus as background distractors. The resulting dataset contains both bag-of-words-equivalent structural adversarial negatives and large-scale heterogeneous background tables, with overall statistics reported in Table 7.

3.4 Human Validation

To assess the quality of Structure-Adv, we randomly sample 300 queries from the 3,000-query test set, covering 300 positive tables and 3,000 corresponding structure-adversarial negatives. We recruit nine annotators with NLP backgrounds. Each example is independently annotated by two annotators, with disagreements resolved by a third adjudicator. The annotation tasks assess query naturalness, positive-table answerability, the correctness and sufficiency of evidence cells, and whether each adversarial negative is genuinely unanswerable.

Table 1 shows that 97.3% of queries are acceptable, 95.3% of positive tables are answerable, and 91.3% of evidence annotations are correct and sufficient. Among the automatically accepted adversarial negatives, 94.0% are verified as unanswerable, while 3.2% remain answerable and 2.8% are judged ambiguous. Raw inter-annotator agreement ranges from 96.0% to 99.0% across four annotation tasks, with Krippendorff’s α between 0.75 and 0.86. These results indicate that the automatic construction pipeline achieves high quality while retaining a small amount of residual annotation noise.

Human-validation item	Result
Query acceptability	97.3%
Positive-table answerability	95.3%
Correct and sufficient evidence	91.3%
Adversarial-negative unanswerability	94.0%
Adversarial negatives that remained answerable	3.2%
Ambiguous adversarial negatives	2.8%
Raw inter-annotator agreement across tasks	96.0%–99.0%
Krippendorff’s α across tasks	0.75–0.86

Table 1: Human-validation results on 300 randomly sampled Structure-Adv queries and their corresponding tables.

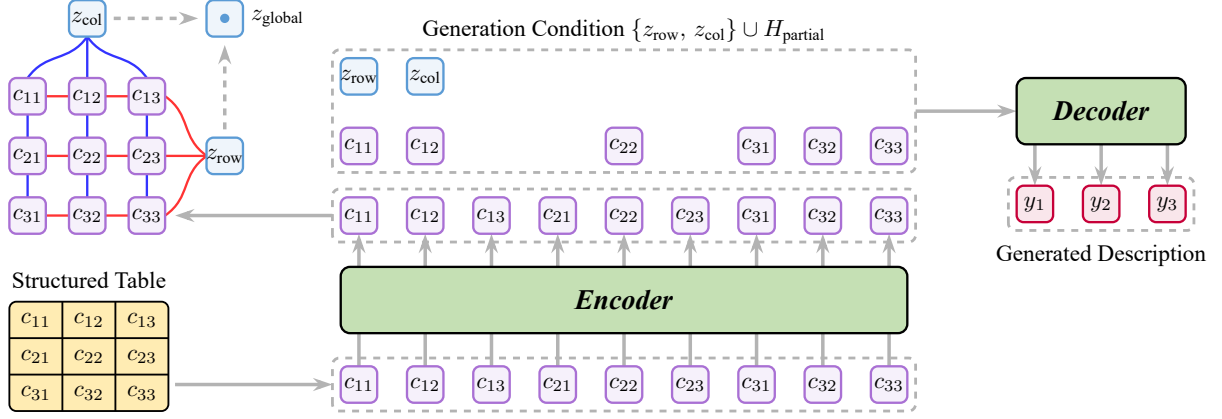


Figure 3: Overall architecture of GSR-Table. The encoder produces multi-view structural representations, while dynamic masking forms H_{partial} as a generative bottleneck. The decoder reconstructs the target description from $\{z_{\text{row}}, z_{\text{col}}\} \cup H_{\text{partial}}$, encouraging the pooled representations to retain row-column binding cues.

4 Methodology

To alleviate bag-of-words degradation in table retrieval, we propose **GSR-Table**, a generative structural regularization framework. The framework contains three key components: multi-view structural feature aggregation, sparse routing with structural compensation, and tri-objective joint optimization. The overall architecture of GSR-Table is illustrated in Figure 3.

4.1 Multi-view Structural Feature Aggregation

Given an input table T , the encoder maps its physical grid structure into the latent space with two-dimensional positional encodings. Let the resulting fine-grained token representations be: $H = \{h_1, h_2, \dots, h_n\}$. Conventional mean pooling or CLS pooling (Devlin et al., 2019) may discard row-column dependencies during representation compression. Therefore, we introduce decoupled attention pooling to extract structural anchors from the row and column dimensions:

$$z_{\text{row}} = \text{AttnPool}_{\text{row}}(H), \quad (4)$$

$$z_{\text{col}} = \text{AttnPool}_{\text{col}}(H). \quad (5)$$

Here, z_{row} captures row-level entity or record information, while z_{col} captures column-level schema or attribute information. These two representations provide complementary structural views for subsequent generative compensation. We then fuse the row and column representations through a nonlinear projection to obtain the global retrieval vector:

$$z_{\text{global}} = \text{MLP}([z_{\text{row}}; z_{\text{col}}]). \quad (6)$$

Thus, z_{global} is derived from decoupled row and column views rather than structure-agnostic global compression, allowing semantic information to be combined with row- and column-sensitive cues.

4.2 Sparse Routing and Generative Structural Compensation

Introducing structural information only on the encoder side does not guarantee that the final retrieval vectors retain structure-related information. If the decoder can access the complete encoder outputs, it may complete generation through shallow lexical matching, leaving z_{row} and z_{col} weakly supervised. To avoid this shortcut, we design a generative information bottleneck.

Specifically, before passing the encoder outputs to the decoder, we apply dynamic masking to table cell representations. Given the encoder output H , the masking function randomly or strategically masks a subset of informative cells, producing a partially visible representation:

$$H_{\text{partial}} = \text{Mask}(H). \quad (7)$$

During decoding, the decoder performs cross-attention over a joint condition set. This set consists of both the multi-view structural representations and the partially visible local representations:

$$\mathcal{C} = \{z_{\text{row}}, z_{\text{col}}\} \cup H_{\text{partial}}. \quad (8)$$

Since H_{partial} contains information gaps, the decoder has less opportunity to rely solely on local token stitching. To reconstruct the target text, it is encouraged to use z_{row} and z_{col} to recover information associated with the masked cells. We refer to this process as *structural compensation*.

Through generation, this process introduces direct back-propagation signals. These signals encourage the row and column representations to retain table content and complementary structure-related cues rather than merely lexical semantics.

Therefore, dynamic masking serves not as a simple data perturbation, but as a sparse routing mechanism for constructing a generative bottleneck. By restricting access to complete local evidence, it transfers reconstruction pressure to the multi-view structural vectors, thereby improving the structure-awareness of the final retrieval representations.

4.3 Tri-objective Joint Optimization

GSR-Table is trained end-to-end with contrastive learning, generative reconstruction, and multi-view orthogonal regularization. The overall objective is to preserve retrieval efficiency while improving the structural robustness of dense table representations.

First, we use the global representation z_{global} as the final retrieval vector of the table, and denote the query representation as q . Given one positive table and a set of negative tables, the contrastive learning loss is defined as:

$$\mathcal{L}_{\text{CL}} = -\log \frac{\exp(\text{sim}(q, z_{\text{global}}^+)/\tau)}{\sum_i \exp(\text{sim}(q, z_{\text{global}}^i)/\tau)}, \quad (9)$$

where $\text{sim}(\cdot, \cdot)$ denotes the similarity function and τ is the temperature coefficient. This loss aligns queries and relevant tables in the retrieval space.

Second, to provide structure-related training signals to the retrieval vectors, we introduce a generative constraint loss. Under the joint condition $\mathcal{C}$, the decoder reconstructs the target textual description $Y = \{y_1, \dots, y_m\}$ with teacher forcing:

$$\mathcal{L}_{\text{Gen}} = -\sum_{t=1}^m \log P(y_t | y_{<t}, \mathcal{C}). \quad (10)$$

Through masked generative reconstruction, this encourages the model to compensate for the missing information using row-column representations.

Finally, to prevent row-level and column-level representations from becoming homogeneous during training, we introduce an orthogonal constraint:

$$\mathcal{L}_{\text{Ortho}} = \frac{(z_{\text{row}}^\top z_{\text{col}})^2}{\|z_{\text{row}}\|^2 \|z_{\text{col}}\|^2}. \quad (11)$$

By minimizing their squared cosine similarity, this constraint encourages z_{row} and z_{col} to remain complementary in the latent space.

The final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CL}} + \lambda_1 \mathcal{L}_{\text{Gen}} + \lambda_2 \mathcal{L}_{\text{Ortho}}, \quad (12)$$

where λ_1 and λ_2 control the weights of generative structural regularization and orthogonal regularization, respectively. With this tri-objective optimization scheme, GSR-Table maintains efficient dense retrieval. It also encourages the final retrieval vectors to retain complementary row-column binding cues, thereby alleviating bag-of-words degradation.

5 Experiments

5.1 Datasets and Baselines

We conduct experiments on two standard table retrieval benchmarks, OTT-QA (Chen et al., 2020) and NQ-Tables (Herzig et al., 2021), as well as our constructed structure-adversarial dataset. The standard benchmarks are used to evaluate general open-domain table retrieval performance. The structure-adversarial dataset further tests model robustness against lexically similar but structurally invalid hard negatives. Detailed dataset descriptions and statistics are provided in Appendix A.

To evaluate GSR-Table, we compare it with representative sparse and dense retrievers. For sparse retrieval, we include BM25 (Robertson and Walker, 1994) and SPLADE (Formal et al., 2021). For the dense retrieval, we compare with BIBERT (Yates et al., 2021), DPR (Wang et al., 2022), TAPAS (Herzig et al., 2020), DTR (Herzig et al., 2021), and SSDR (Jin et al., 2023). All models are evaluated under the same evaluation protocols. The evaluation metrics include MRR@10, Recall@k, and NDCG@k (Voorhees and Tice, 2000; Järvelin and Kekäläinen, 2002). Detailed descriptions of the baselines are provided in Appendix B.

5.2 Implementation Details

GSR-Table adopts an encoder-decoder transformer architecture. We construct a separate training corpus of 922K high-quality tables from the WDC corpus (Lehmberg et al., 2016). During corpus construction, we exclude any table whose serialized content exactly matches one of the 63,209 source tables used to construct Structure-Adv. We then use Gemma-4-31B (Google DeepMind, 2026b) to generate a natural-language query and a table description for each table. To enhance the model’s structural compensation ability, we apply cell-level random masking to the encoder hidden states with

	NQ-TABLES					OTT-QA				
	NDCG@5	NDCG@10	R@1	R@10	R@50	NDCG@5	NDCG@10	R@1	R@10	R@50
BM25	25.52	27.12	18.49	36.94	52.61	35.09	37.45	23.98	51.94	69.11
SPLADE	58.46	61.27	38.43	83.13	95.25	76.31	77.55	64.45	89.39	95.89
BIBERT	61.73	63.16	44.49	84.87	94.16	70.28	72.49	<u>57.05</u>	86.45	94.26
DPR	63.05	66.13	45.32	<u>85.84</u>	<u>95.44</u>	67.92	70.00	53.43	85.95	93.22
TAPAS	61.08	63.89	43.79	83.49	95.10	<u>70.44</u>	<u>72.72</u>	56.05	86.72	94.49
DTR	49.13	52.70	31.72	74.63	91.10	57.23	59.30	42.14	75.93	87.58
SSDR _{im}	62.31	65.02	<u>45.47</u>	84.00	95.05	69.81	71.76	56.96	86.22	93.55
GSR-Table	<u>62.58</u>	<u>65.89</u>	46.84	86.63	96.46	69.55	72.15	56.42	<u>87.42</u>	<u>95.60</u>

Table 2: Main retrieval results on two standard table retrieval benchmarks, NQ-TABLES and OTT-QA. Bold numbers indicate the best results, and underlined numbers indicate the second-best results.

a masking probability of 0.15. For retrieval training, the source table corresponding to each query is treated as the positive instance, while other tables in the same mini-batch are used as in-batch negatives. During inference, we discard the decoder and use only the encoder-side table representation for dense retrieval. More implementation and training details are provided in Appendix C. A one-variable-at-a-time sensitivity analysis of λ_1 and λ_2 is reported in Appendix D.

5.3 Main Results on Standard Table Retrieval Benchmarks

Table 2 reports the main results of different methods on two standard table retrieval benchmarks, NQ-TABLES and OTT-QA. Overall, GSR-Table achieves competitive retrieval performance on both datasets, with particularly strong results on recall-oriented metrics. These results suggest that the proposed generative structural regularization is effective for target-table retrieval.

On NQ-TABLES, GSR-Table achieves 46.84, 86.63, and 96.46 on R@1, R@10, and R@50, respectively, outperforming all compared methods. In particular, its R@1 score improves over the strongest baseline SSDR_{im} by 1.37 percentage points, indicating that GSR-Table can more accurately rank the relevant table at the top of the retrieved list. GSR-Table also obtains the best results on R@10 and R@50, showing that it can consistently cover the target table when a larger candidate set is considered. Although GSR-Table is slightly lower than DPR on NDCG@5 and NDCG@10, its stronger overall recall demonstrates its practical effectiveness in open-domain table retrieval.

On OTT-QA, SPLADE achieves strong overall performance as a sparse retrieval model, suggesting that lexical matching signals remain highly ef-

fective on this dataset. Nevertheless, GSR-Table remains competitive among dense retrieval methods on this challenging benchmark. In particular, it achieves 87.42 on R@10 and 95.60 on R@50, outperforming all dense retrieval baselines. These results suggest that GSR-Table remains effective among dense retrievers, particularly in terms of target-table recall over larger candidate sets.

	MRR@10	R@1	R@5	R@10
BM25	7.72	2.84	16.65	38.47
SPLADE	22.36	7.90	38.24	66.85
BIBERT	26.94	12.27	44.72	69.36
DPR	29.15	14.43	47.86	72.52
TAPAS	34.87	19.63	58.31	81.44
DTR	31.68	17.47	54.26	79.19
SSDR _{im}	<u>38.45</u>	<u>22.73</u>	<u>62.58</u>	<u>83.76</u>
GSR-Table	43.82	28.63	69.39	88.27

Table 3: Robustness results under structure-adversarial evaluation. The best and second-best results are marked in bold and underlined, respectively.

5.4 Robustness under Structure-Adversarial Evaluation

To evaluate model robustness under structure-adversarial scenarios, we conduct experiments on the structure-adversarial dataset constructed in Section 3. Specifically, each query is associated with one positive table, ten structure-adversarial negatives, and a large set of background tables. We use MRR@10 and Recall@k as the main evaluation metrics, where $k \in \{1, 5, 10\}$.

The results are shown in Table 3. Sparse retrieval methods perform poorly in this setting. SPLADE achieves 22.36 on MRR@10 and 7.90 on R@1, suggesting that lexical matching struggles with the joint interference from structure-adversarial negatives and background tables. Standard dense re-

trievers improve over sparse methods, but BIBERT and DPR still obtain only 12.27 and 14.43 on R@1, respectively. In contrast, TAPAS, DTR, and SSDR_{im} achieve better results, indicating that table modeling and fine-grained representations improve robustness in this retrieval setting.

GSR-Table achieves the best performance across all metrics, with 43.82, 28.63, 69.39, and 88.27 on MRR@10, R@1, R@5, and R@10, respectively. Compared with the strongest baseline SSDR_{im} , GSR-Table improves by 5.37, 5.90, 6.81, and 4.51 percentage points, respectively. These results show that GSR-Table more reliably retrieves the target table under structure-adversarial negatives and large-scale background distractors. We further report the controlled 11-way adversarial ranking results in Appendix E.2 to directly analyze each model’s ability to distinguish BoW-equivalent structural negatives.

To examine whether the residual annotation noise affects model comparisons, we re-evaluate all models on the adjudicated 300-query subset. GSR-Table remains the best-performing model across all metrics, outperforming SSDR_{im} by 5.49 points in MRR@10 and 5.67 points in R@1, while the overall model ranking remains unchanged after this correction (Appendix E.1).

5.5 Training-Matched Controls

To isolate architectural effects from large-scale synthetic supervision, we train four control models from scratch on the same 922K query–table pairs. All controls use the same encoder depth and hidden size as GSR-Table. The *Linearized Dual Encoder* uses one-dimensional header–cell-value sequences, while the *2D Dual Encoder* adds row and column positional embeddings but retains global pooling; both use only contrastive learning. The *SSDR_{im}-style Retriever* uses multi-vector MaxSim interaction. The *Global-Gen* retains the encoder–decoder architecture and description-generation objective, but uses global pooling and full encoder states without row/column representations, dynamic masking. We evaluate all controls on Structure-Adv using the same candidate pool and evaluation protocol.

Table 4 shows that GSR-Table outperforms all training-matched controls on the Structure-Adv. Global-Gen is the strongest control, but global description generation alone does not reproduce the full model’s gains. Nevertheless, GSR-Table improves over Global-Gen by 2.41, 3.20, 3.62, and 3.19 points on MRR@10, R@1, R@5, and R@10, respectively. Because all models receive

Model	MRR@10	R@1	R@5	R@10
Linearized Dual Encoder	30.82	16.73	52.91	76.84
2D Dual Encoder	34.91	20.00	58.26	79.85
SSDR_{im} -style Retriever	39.64	23.90	64.38	84.91
Global-Gen	<u>41.41</u>	<u>25.43</u>	<u>65.77</u>	<u>85.08</u>
GSR-Table	43.82	28.63	69.39	88.27

Table 4: Training-matched results on Structure-Adv. All models use the same 922K synthetic training pairs, encoder depth, and hidden size.

the same synthetic supervision, these results indicate that GSR-Table’s gains cannot be attributed solely to the additional training data or description-generation objective.

5.6 Analysis of Bag-of-Words Degradation

To analyze bag-of-words degradation, we construct BoW-enhanced adversarial negatives. Specifically, we inject query tokens and evidence-path tokens into structure-adversarial negatives sampled from our test set. These negatives have higher lexical overlap with the query but remain unanswerable, directly testing whether a model is misled by surface lexical matching. We report ASR_{BoW} , Pair-Acc, and Margin; the test set construction details and metric definitions are provided in Appendix F.

The results are shown in Table 6. Overall, existing retrieval models exhibit clear degradation under BoW-enhanced negatives. BM25 and SPLADE obtain ASR_{BoW} scores of 97.00 and 84.67, respectively. This shows that when negatives are injected with more query-related tokens, sparse retrieval methods tend to misidentify them as more relevant tables. These results suggest that term-matching retrievers are easily misled by surface lexical overlap and struggle to identify structurally valid evidence.

Although dense retrievers improve over sparse methods, they still suffer from severe bag-of-words degradation. BIBERT and DPR obtain ASR_{BoW} scores of 70.67 and 66.33, respectively, meaning that in more than half of the cases, they assign higher relevance scores to the BoW-enhanced negatives. TAPAS, DTR, and SSDR_{im} perform better, suggesting that table modeling and fine-grained representations partly alleviate this issue. However, even the strongest baseline, SSDR_{im} , reaches an ASR_{BoW} of 39.00, indicating that existing methods are still affected by lexical shortcuts.

In contrast, GSR-Table is the most robust in this experiment, with an ASR_{BoW} of only 20.33 and a Margin of 0.133. Compared with the strongest baseline SSDR_{im} , GSR-Table reduces ASR_{BoW} by

	NQ-TABLES					Structure-Adv			
	NDCG@5	NDCG@10	R@1	R@10	R@50	MRR@10	R@1	R@5	R@10
GSR-Table	62.58	65.89	46.84	86.63	96.46	43.82	28.63	69.39	88.27
w/o $\mathcal{L}_{\text{Gen}}$	60.34	63.74	44.21	84.58	94.72	34.47	19.87	58.32	79.64
w/o Dynamic Masking	60.72	64.12	44.96	85.03	95.07	37.62	22.73	61.45	82.16
w/o $\mathcal{L}_{\text{Ortho}}$	61.76	65.12	45.83	86.01	95.93	41.68	26.70	67.14	86.76
w/o Row/Column Views	61.49	64.93	45.54	85.55	95.40	41.52	26.17	66.74	86.17

Table 5: Ablation results on NQ-TABLES and the structure-adversarial benchmark.

	ASR _{BoW} ↓	Pair-Acc ↑	Margin ↑
BM25	97.00	3.00	-0.312
SPLADE	84.67	15.33	-0.221
BIBERT	70.67	29.33	-0.106
DPR	66.33	33.67	-0.084
TAPAS	45.67	54.33	0.028
DTR	53.33	46.67	-0.046
SSDR _{im}	<u>39.00</u>	<u>61.00</u>	<u>0.071</u>
GSR-Table	20.33	79.67	0.133

Table 6: Analysis of bag-of-words degradation under BoW-enhanced adversarial negatives. Bold and underline indicate the best and second-best results.

18.67 percentage points. This suggests that GSR-Table better resists lexically enhanced negatives and favors structurally valid positive tables. Overall, this experiment reveals that existing table retrieval models commonly suffer from bag-of-words degradation, while GSR-Table effectively mitigates lexical shortcut exploitation.

5.7 Ablation Study

We conduct ablation studies to verify the effectiveness of each key component in GSR-Table. Table 5 reports the results of different model variants on NQ-TABLES and Structure-Adv. Overall, removing any component leads to performance degradation, with more pronounced drops on Structure-Adv, indicating that these modules are important for improving structural robustness.

First, removing generative reconstruction loss causes the largest performance drop. On Structure-Adv, MRR@10 decreases from 43.82 to 34.47, and R@1 decreases from 28.63 to 19.87. This suggests that generative structural regularization provides effective structural supervision for retrieval representations. Second, when the dynamic masking mechanism is removed, MRR@10 and R@1 on Structure-Adv drop to 37.62 and 22.73, respectively. This result indicates that if the decoder can directly access the full encoder outputs, the generative task may degenerate into local token copying,

thereby weakening the information bottleneck.

In addition, removing the orthogonal constraint also leads to performance degradation, although the drop is relatively smaller. This suggests that maintaining the complementarity between z_{row} and z_{col} helps mitigate the homogenization of row- and column-level representations. Finally, the *w/o Row/Column Views* variant replaces z_{row} and z_{col} with a single global token, which serves as the final retrieval representation. Its lower performance indicates that explicit row/column views provide structural cues that cannot be fully recovered by a single global token, even with masked generative supervision.

6 Conclusion

In this paper, we studied bag-of-words degradation in table retrieval and introduced a structure-adversarial benchmark to evaluate models’ sensitivity to row-column binding perturbations. The benchmark generates lexically equivalent but structurally invalid hard negatives through evidence-cell anchoring and Homologous Structural Swapping. To mitigate this issue, we proposed GSR-Table, a generative structural regularization framework for table retrieval. GSR-Table models row- and column-level structural anchors with orthogonal attention pooling, and uses dynamic masking to encourage structural compensation during reconstruction. Experimental results show that GSR-Table remains competitive on standard table retrieval benchmarks and substantially improves robustness under structure-adversarial settings.

Limitations

This work has several limitations. First, our structure-adversarial benchmark mainly targets relational-style tables with clear row-column semantics. Its applicability to irregular tables, nested headers, multi-table reasoning, and visually complex table documents remains to be further studied. Second,

benchmark construction relies on LLM-generated queries, evidence annotations, and answerability verification. Although we apply consistency filtering, annotation noise and ambiguous cases may still exist.

Third, GSR-Table introduces decoder-side generative regularization during training, which requires each table to be paired with a high-quality textual description. Such supervision may not always be readily available in real-world table retrieval scenarios, especially for large-scale or domain-specific table corpora. Fourth, although we remove exact matches between the 922K training corpus and Structure-Adv source tables, this filtering does not rule out near-duplicate tables, similar schemas, or shared domain patterns within WDC.

Ethical Considerations

This work uses public table corpora and does not involve private user data. However, the use of LLMs for query generation and answerability verification may introduce factual errors, biases, or annotation artifacts. We mitigate these risks through filtering and closed-loop verification, and will release the construction pipeline for auditing. The proposed adversarial benchmark is intended for diagnosing structural weaknesses in table retrievers. Nevertheless, similar techniques could be misused to create lexically plausible but structurally misleading tables. We encourage the use of our benchmark for robustness evaluation and defensive research. In high-stakes domains, retrieved tables should still be verified by humans or domain experts, since improved robustness does not guarantee factual correctness in all real-world scenarios.

Acknowledgements

This work was supported by the Project 62276178 under the National Natural Science Foundation of China, the Key Project 23KJA520012 under the Natural Science Foundation of Jiangsu Higher Education Institutions, Project 24BTQ002 under the National Social Science Foundation of China, the project 22YJCZH091 of Humanities and Social Science Fund of Ministry of Education and the Priority Academic Program Development of Jiangsu Higher Education Institutions. They are also with Jiangsu Key Lab of Language Computing, Suzhou, 215021, P. R. China.

References

- Iñigo Alonso, Eneko Agirre, and Mirella Lapata. 2024. [PixT3: Pixel-based table-to-text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6721–6736, Bangkok, Thailand. Association for Computational Linguistics.
- Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. [The fact extraction and VERification over unstructured and structured information \(FEVEROUS\) shared task](#). In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.
- Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. 2015. Tabel: Entity linking in web tables. In *International Semantic Web Conference*, pages 425–441. Springer.
- Michael J Cafarella, Alon Y Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. Webtables: Exploring the power of tables on the web. *Proc. VLDB Endow.*, 1(1):538–549.
- Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.
- Google DeepMind. 2026a. Gemini 3.1 pro. <https://deepmind.google/models/gemini/pro/>. Accessed: May 11, 2026.
- Google DeepMind. 2026b. Gemma 4. <https://deepmind.google/models/gemma/gemma-4/>. Accessed: May 11, 2026.
- Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. [Open domain question answering over tables via dense retrieval](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

- Linguistics: Human Language Technologies*, pages 512–519, Online. Association for Computational Linguistics.
- Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. [TaPas: Weakly supervised table parsing via pre-training](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.
- Robin Jia and Percy Liang. 2017. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Nengzheng Jin, Dongfang Li, Junying Chen, Joanna Siebert, and Qingcai Chen. 2023. [Enhancing open-domain table question answering via syntax- and structure-aware dense retrieval](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Nusa Dua, Bali. Association for Computational Linguistics.
- Mihir Kale and Abhinav Rastogi. 2020. [Text-to-text pre-training for data-to-text tasks](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.
- Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. 2016. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th international conference companion on world wide web*, pages 75–76.
- Da Li, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2025. [Tailoring table retrieval from a field-aware hybrid matching perspective](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27693–27704, Suzhou, China. Association for Computational Linguistics.
- WeiJie Liu, Yibin Zheng, and Fang Kong. 2025. [QuASAR: A question-driven structure-aware approach for table-to-text generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26798–26812, Vienna, Austria. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- OpenAI. 2025. GPT-5 mini model. <https://developers.openai.com/api/docs/models/gpt-5-mini>. Accessed: May 11, 2026.
- Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR’94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.
- Adarsh Singh, Kushal Raj Bhandari, Jianxi Gao, Soham Dan, and Vivek Gupta. 2025. Craft: Training-free cascaded retrieval for tabular qa. *arXiv preprint arXiv:2505.14984*.
- Ellen M. Voorhees and Dawn M. Tice. 2000. [The TREC-8 question answering track](#). In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00)*, Athens, Greece. European Language Resources Association (ELRA).
- Zhiruo Wang, Zhengbao Jiang, Eric Nyberg, and Graham Neubig. 2022. Table retrieval may not necessitate table-specific model design. In *Proceedings of the workshop on structured and unstructured knowledge integration (SUKI)*, pages 36–46.
- Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.
- Shuo Zhang and Krisztian Balog. 2018. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 world wide web conference*, pages 1553–1562.

A Dataset Details

This section provides detailed descriptions of the datasets used in our experiments. Detailed statistics for all datasets are reported in Table 7.

NQ-TABLES. NQ-TABLES is an open-domain table question answering dataset derived from Natural Questions. Models must retrieve relevant tables from a large-scale Wikipedia table corpus before answer extraction. Each query is typically paired with a reference table, making the dataset suitable for evaluating open-domain target-table retrieval.

OTT-QA. OTT-QA is an open-domain table-and-text QA dataset built upon HybridQA, with decontextualized questions for retrieval. The original task retrieves table segments and text passages for multi-hop reasoning. As we focus on table retrieval, we retain only table candidates and exclude textual evidence from our experiments.

		NQ-TABLES		OTT-QA		Structure-Adv	
		Train	Test	Train	Test	Train	Test
Query	Count	9,594	919	41,469	2,214	60,209	3,000
	Avg. # Words.	8.94	8.90	21.79	22.82	14.63	14.39
Table	Count	169,898	169,898	419,183	419,183	413,209	413,209
	Avg. # Row.	10.70	10.70	12.90	12.90	7.83	7.83
	Avg. # Col.	6.10	6.10	4.80	4.80	6.44	6.44
# Relevant Tables per Query		1.00	1.05	1.00	1.00	1.00	1.00

Table 7: Statistics of the datasets used in our experiments, including query counts, table corpus sizes, average query lengths, average table dimensions, and the number of relevant tables per query.

Structure-Adversarial Dataset. Structure-Adv is constructed in Section 3 to evaluate structural robustness in table retrieval. It tests whether models capture entity-attribute-value bindings beyond shallow lexical matching. For each query, it provides lexically similar hard negatives that break the evidence path and cannot support the query.

For Structure-Adv, the shared retrieval corpus contains all 63,209 source tables, 30,000 retained HSS negatives, and 320,000 background distractors, for a total of 413,209 tables. Consequently, the training and test columns in Table 7 have different query counts but the same table-corpus size.

A.1 Structure-Adv Construction and Validation

Construction Statistics. After consistency verification, we retain 63,209 valid query-answer-evidence triples from the 68,000 generated candidates. We then adopt a test-first dynamic split: for each screened source example, we generate HSS candidates, reject those that remain answerable, and include the example in the test set only when at least ten valid negatives are available. Screening continues until 3,000 test queries are collected, with exactly ten negatives retained per query; the remaining 60,209 queries form the training set. In total, 55,717 HSS candidates are generated for 3,880 screened examples, of which 34,894 pass verification and 30,000 are retained in the final test set. Table 10 summarizes query-answer-evidence generation and dataset splitting, while Table 11 reports HSS candidate generation and verification in detail.

Human Validation. Human validation covers 300 randomly sampled test queries, their 300 positive tables, and the corresponding 3,000 HSS negatives. Each item is independently evaluated by two annotators, with disagreements resolved by a third adjudicator. For positive instances, the evaluation

Query-quality category	Percentage
Acceptable	97.3%
Unacceptable	0.0%
Ambiguous	2.7%

Table 8: Distribution of query quality in the human evaluation of sampled test queries.

Evidence-annotation category	Percentage
Correct and sufficient	91.3%
Incorrect or insufficient	6.7%
Ambiguous	2.0%

Table 9: Validity of automatically generated evidence-cell annotations in the human evaluation.

covers query naturalness, table answerability, and the correctness and sufficiency of the highlighted evidence cells. For negative instances, annotators classify each perturbed table as answerable, unanswerable, or ambiguous. Tables 8–13 report the category-level distributions, raw agreement, Krippendorff’s α , and adjudication rates.

B Baseline Descriptions

Sparse Retrieval. BM25 (Robertson and Walker, 1994) is a term-matching method that measures query-table relevance based on term frequency, inverse document frequency, and document-length normalization. SPLADE (Formal et al., 2021) is a neural sparse retriever built on pretrained language models. It learns the sparse representations through lexical expansion and sparsity regularization, thereby alleviating the vocabulary mismatch problem in traditional sparse retrieval. These two methods represent conventional term matching and neural sparse retrieval, respectively.

Dense Retrieval. BIBERT (Yates et al., 2021) is a basic dual-encoder retriever that linearizes tables into one-dimensional text and separately encodes

Construction stage	Number	Percentage
Generated query–answer–evidence triples	68,000	—
Rejected by consistency verification	4,791	7.05%
Valid triples retained	63,209	92.95%
Samples screened for test-set construction	3,880	6.14% of valid
Samples meeting the ten-negative requirement	3,000	77.32% of screened
Reassigned to training: fewer than ten valid negatives	880	22.68% of screened
Remaining samples assigned to training	59,329	93.86% of valid
Final test queries	3,000	4.75% of valid
Final training queries	60,209	95.25% of valid

Table 10: Stage-wise statistics of query–answer–evidence generation and the test-first dynamic splitting procedure.

Statistic	Number	Percentage
Source examples screened	3,880	—
HSS candidates generated	55,717	100.00%
Rejected as still answerable	20,823	37.37%
Verified as structurally invalid and unanswerable	34,894	62.63%
Retained in the final test set	30,000	53.84%
Not retained: source examples had fewer than ten valid negatives	4,894	8.78%
Avg. candidates per screened example	14.36	—
Final negatives per test query	10	—

Table 11: Stage-wise statistics of HSS candidate generation and answerability verification. Percentages use the 55,717 generated candidates as the denominator.

queries and tables for vector similarity computation. DPR (Wang et al., 2022) follows a similar dual-encoder architecture, but strengthens contrastive learning with hard negatives, making it a strong baseline in open-domain retrieval. TAPAS (Herzig et al., 2020) is a table-understanding pretrained model that explicitly incorporates structural signals such as row indices, column indices, and numerical features. DTR (Herzig et al., 2021) is a contrastive dual-encoder retriever that incorporates row and column features to improve structural modeling for table retrieval. SSDR (Jin et al., 2023) adopts a syntax- and structure-aware multi-vector late-interaction mechanism, enabling fine-grained matching between queries and structural table units for query–table alignment.

C Implementation and Training Details

The encoder and decoder of GSR-Table each consist of 12 Transformer layers, with the hidden size set to 768. All model parameters are randomly initialized, and no pretrained language model is used. We train the model using the AdamW optimizer with a learning rate of 1×10^{-4} . The warmup ratio is set to 0.1, followed by a linear learning rate decay schedule. The overall training objec-

tive consists of the query-table contrastive loss, the generative structural regularization loss, and the orthogonality constraint. The loss weights of the two auxiliary objectives are set to $\lambda_1 = 0.3$ and $\lambda_2 = 0.05$, respectively.

The retrieval score is computed as the similarity between the query representation and the global table representation, and the model is optimized with a contrastive learning objective. For random masking, all token representations belonging to the selected cells are masked in the encoder hidden states. The masking operation is applied before decoder cross-attention.

During synthetic data construction, we further filter out samples with invalid formats, overly short contents, or duplicated generations to ensure training data quality. The model is trained for 30 epochs on 4 NVIDIA RTX 4090 GPUs, taking approximately 13 hours in total. Unless otherwise specified, all experiments use the same training configuration.

D Hyperparameter Sensitivity

We conduct a sensitivity analysis for the two auxiliary loss weights in Equation (12). Specifically, we vary $\lambda_1 \in \{0.2, 0.3, 0.4, 0.5\}$ while fixing

Object	Answerable	Unanswerable	Ambiguous
Positive table	95.3%	1.3%	3.3%
Accepted adversarial negative	3.2%	94.0%	2.8%

Table 12: Human judgments of positive- and adversarial-negative-table answerability. Percentages may not sum to 100% because of independent rounding.

Annotation task	Sample size	Raw agreement	Krippendorff’s α	Adjudication rate
Query naturalness	300	99.0%	0.84	1.0%
Positive answerability	300	98.7%	0.86	1.3%
Evidence correctness	300	96.0%	0.75	4.0%
Negative answerability	3,000	97.7%	0.80	2.3%

Table 13: Inter-annotator agreement and adjudication rates across the four human-annotation tasks.

$\lambda_2 = 0.05$, and then vary $\lambda_2 \in \{0.05, 0.10, 0.15\}$ while fixing $\lambda_1 = 0.3$. The candidate values are selected according to training stability, the convergence of the individual loss terms, and validation performance. We report NDCG@5, NDCG@10, R@1, R@10, and R@50 on NQ-TABLES, together with MRR@10, R@1, R@5, and R@10 on Structure-Adv. As shown in Table 14, GSR-Table remains relatively stable across the evaluated ranges of both loss weights. The selected setting, $\lambda_1 = 0.3$ and $\lambda_2 = 0.05$, provides the best overall performance on both datasets. Increasing either weight beyond the selected value gradually reduces performance, suggesting that overly strong auxiliary regularization may interfere with the primary contrastive retrieval objective.

E Additional Robustness Results

E.1 Evaluation on the Human-Verified Subset

To test whether residual annotation noise affects model comparisons, we construct Human-Gold from the 300-query audit described in Section 3. Using only the adjudicated annotations, we correct ambiguous queries, inaccurate evidence annotations, and invalid adversarial negatives without consulting model outputs, and re-evaluate all models under the same open-domain setting.

As shown in Table 15, GSR-Table remains the strongest model across all four metrics and outperforms SSDR_{im} by 5.49, 5.67, 7.03, and 4.41 percentage points on MRR@10, R@1, R@5, and R@10, respectively. Compared with the full automatically constructed test set in Table 3, performance changes only slightly and the relative ordering remains unchanged. This consistency indicates that GSR-Table’s advantage is robust to the annota-

tion noise identified by the human audit.

E.2 Controlled 11-way Adversarial Ranking Results

Unlike the open-domain setting in Section 5.4, the controlled 11-way adversarial ranking does not include large-scale background tables. For each query, the candidate set contains one positive table and ten structure-adversarial negatives. This setting directly tests whether a retriever can rank the structurally valid table above BoW-equivalent but structurally invalid negatives.

Table 16 reports the detailed results. BM25 relies on random tie-breaking because it cannot distinguish BoW-equivalent candidates, while SPLADE only brings limited improvement. Dense retrievers perform better overall, but BIBERT and DPR remain vulnerable to structure-adversarial negatives. TAPAS, DTR, and SSDR_{im} achieve stronger results, suggesting that table structure modeling and fine-grained interaction are helpful.

GSR-Table achieves the best performance in this setting. Compared with SSDR_{im}, it improves MRR@10, R@1, and R@5 by 8.75, 12.24, and 9.28 percentage points, respectively. This shows that GSR-Table can more effectively distinguish structurally valid tables from BoW-equivalent structural negatives.

F Details of Bag-of-Words Degradation Analysis

This section details the construction of our BoW-enhanced adversarial negatives and the evaluation metrics. The experiment tests whether retrievers degenerate into bag-of-words matching by preferring negatives with more query-related tokens.

Varied weight	λ_1	λ_2	NQ-TABLES					Structure-Adv			
			NDCG@5	NDCG@10	R@1	R@10	R@50	MRR@10	R@1	R@5	R@10
λ_1	0.2	0.05	61.64	64.94	45.88	85.75	95.74	42.51	27.37	68.16	87.21
	0.3	0.05	62.58	65.89	46.84	86.63	96.46	43.82	28.63	69.39	88.27
	0.4	0.05	62.32	65.62	46.55	86.52	96.58	43.37	28.13	68.96	87.92
	0.5	0.05	60.12	64.16	45.11	85.03	95.32	42.28	26.30	67.91	86.94
λ_2	0.3	0.05	62.58	65.89	46.84	86.63	96.46	43.82	28.63	69.39	88.27
	0.3	0.10	62.13	65.63	46.29	86.21	95.88	42.41	27.20	68.01	86.95
	0.3	0.15	58.43	62.51	42.68	82.34	93.05	39.83	25.67	66.51	85.54

Table 14: Sensitivity analysis of λ_1 and λ_2 on NQ-TABLES and Structure-Adv. One loss weight is varied while the other is fixed at its default value.

Model	MRR@10	R@1	R@5	R@10
BM25	8.11	3.00	17.37	39.72
SPLADE	23.18	8.33	39.77	68.17
BIBERT	27.10	13.00	45.93	70.30
DPR	29.72	15.33	49.00	73.72
TAPAS	35.44	20.67	59.67	82.93
DTR	32.25	18.33	55.33	80.27
SSDR _{im}	<u>39.42</u>	<u>24.00</u>	<u>64.00</u>	<u>84.93</u>
GSR-Table	44.91	29.67	71.03	89.34

Table 15: Retrieval results on the 300-query Human-Gold subset. The best and second-best results are marked in bold and underlined, respectively.

	MRR@10	R@1	R@5	R@10
BM25	26.63	9.09	45.45	90.91
SPLADE	34.15	14.63	57.27	93.06
BIBERT	41.72	23.33	61.18	94.47
DPR	44.08	25.67	64.31	95.26
TAPAS	50.96	32.83	77.62	<u>98.18</u>
DTR	47.24	29.57	72.36	97.74
SSDR _{im}	<u>53.63</u>	<u>35.93</u>	<u>79.45</u>	98.02
GSR-Table	62.38	48.17	88.73	99.21

Table 16: Controlled 11-way adversarial ranking results. Each query is evaluated against one positive table and ten BoW-equivalent structure-adversarial negatives. Bold and underline indicate the best and second-best results.

Test Set Construction. We randomly sample 300 examples from the structure-adversarial test set constructed in Section 3. For each query, we inject query and evidence-path tokens into non-critical cells of an unanswerable negative, increasing its lexical overlap with the query. The token injection does not restore the original entity-attribute-value bindings, so the BoW-enhanced negative remains unanswerable. This directly tests whether retrievers are misled by surface lexical overlap. Appendix G shows an example construction process.

Evaluation Metrics. Let T_i^+ denote the positive table for the i -th query, $T_{\text{bow},i}^-$ denote the corresponding BoW-enhanced negative table, and $s(q_i, T)$ denote the relevance score assigned by a retriever to a query-table pair. We use three metrics for evaluation.

First, ASR_{BoW} directly measures the proportion of cases where the BoW-enhanced negative receives a higher score than the positive table:

$$\text{ASR}_{\text{BoW}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I} \left[s(q_i, T_{\text{bow},i}^-) > s(q_i, T_i^+) \right].$$

Second, Pair-Acc measures whether the model assigns a higher score to the positive table than to the BoW-enhanced negative. Since this metric is complementary to ASR_{BoW} , we compute it as:

$$\text{PairAcc} = 1 - \text{ASR}_{\text{BoW}}.$$

Finally, since different retrievers produce scores on different scales, we use the normalized score margin $\text{Margin}_{\text{BoW}}$ to measure the relative score gap between the positive and negative tables:

$$\text{Margin} = \frac{1}{N} \sum_{i=1}^N \frac{s(q_i, T_i^+) - s(q_i, T_{\text{bow},i}^-)}{|s(q_i, T_i^+)| + |s(q_i, T_{\text{bow},i}^-)| + \epsilon}.$$

Lower ASR_{BoW} , higher Pair-Acc, and higher $\text{Margin}_{\text{BoW}}$ indicate stronger resistance to bag-of-words shortcuts.

G Example of BoW-enhanced Adversarial Negative Construction

Query: Which event is planned for beginners and takes place in the Kitchen?

Positive Table T^+

Event	Location	Audience	Status	Participants	Sessions	Note
Book Club	Room A	Adults	Popular	24	3	Weekly
Movie Night	Main Hall	Families	Full	46	2	Weekend
Yoga Class	Garden	Seniors	Open	18	4	Morning
Cooking Demo	Kitchen	Beginners	Planned	32	2	Indoor
Art Workshop	Studio	Students	Active	27	3	Hands-on

Structure-Adversarial Negative T^-

Event	Location	Audience	Status	Participants	Sessions	Note
Book Club	Room A	Adults	Planned	24	3	Weekly
Movie Night	Main Hall	Families	Full	46	2	Weekend
Yoga Class	Garden	Beginners	Open	18	4	Morning
Cooking Demo	Kitchen	Seniors	Popular	32	2	Indoor
Art Workshop	Studio	Students	Active	27	3	Hands-on

BoW-enhanced Negative T_{bow}^-

Event	Location	Audience	Status	Participants	Sessions	Note
Book Club	Room A	Adults	Planned	24	3	Planned event
Movie Night	Main Hall	Beginners	Full	46	2	Weekend
Yoga Class	Garden	Beginners	Open	18	4	Morning
Cooking Demo	Kitchen	Seniors	Popular	32	2	Kitchen
Art Workshop	Studio	Students	Active	27	3	Beginners planned

Figure 4: Example of BoW-enhanced adversarial negative construction. The positive table T^+ contains a valid structural binding for the query, while T^- breaks this binding. T_{bow}^- further injects query-related tokens into non-critical cells, increasing lexical overlap without restoring the valid evidence path.