DSS: Dynamic Sparse Selection for Multi-modal Knowledge Graph Completion

1st Yi Zhou
School of Computer Science and Technology
Southwest University of Science and Technology
Mianyang, China
553580053@qq.com

2nd Qingming Zhang*
School of Computer Science and Technology
Southwest University of Science and Technology
Mianyang, China
qmzhang@swust.edu.cn

***Abstract*—Multi-modal Knowledge Graph Completion (MMKGC) aims to uncover hidden knowledge by fusing multi-modal and structured information. Existing fusion methods use attention, weighting, or mixture-of-experts strategies. However, they treat all feature segments equally at the fine-grained level. This equal treatment prevents effective filtering of image background noise and textual redundancy. Consequently, discriminative features are weakened. To address this, this paper introduces a sparse mechanism into the model, proposing a Dynamic Sparse Selection (DSS) model. Its core is the Sparse-Global Fusion (S-GF) method, which utilizes Sparse Prompt Attention (SPA) and Global Dense Attention (GDA) in parallel. The SPA branch uses a lightweight controller to learn attention masks. It focuses on the most discriminative local segments. Meanwhile, the GDA branch captures global context to preserve semantic coherence. The two are adaptively fused via a gating mechanism. Additionally, we propose Sparse Gated Contrastive Learning (SGCL). DSS proposes the Sparse Gated Selection Mechanism (SGSM) to dynamically construct tailored anchor representations for entities. Optimization is performed using the Multi-view Focused Adaptive Contrastive Loss (MFAC Loss) with progressive focusing capability, thereby enhancing both the discriminative power and alignment robustness of entity representations. Experiments on DB15K, MKG-W, and MKG-Y show that DSS improves the MRR metric by 6.48%, 2.75%, and 0.97%, respectively, compared to existing MMKGC models, demonstrating the effectiveness of the introduced sparse mechanism in focusing on critical information and suppressing multi-modal noise.**

***Index Terms*—knowledge graph, knowledge graph completion, multi-modal knowledge graph completion, contrastive learning, link prediction**

I. Introduction

Multi-modal Knowledge Graphs (MMKGs) [1], [2] connect text, images, and other multi-modal information to the structured triplets of traditional Knowledge Graphs (KGs). This enhances the richness of knowledge representation and finds wide application in areas like recommendation systems [3] and large language models [4]. However, MMKGs, like single-modal KGs, commonly face incompleteness. The Multi-modal Knowledge Graph Completion (MMKGC) task [1], [2] aims to leverage existing multi-modal knowledge to build more comprehensive knowledge representations. It predicts missing world knowledge in MMKGs, thereby mitigating this limitation.

Despite progress in MMKGC research, existing methods still face two unresolved key problems in achieving fine-grained, precise multi-modal fusion and alignment.

Insufficient Dynamic Information Filtering in Fine-grained Fusion. Current mainstream MMKGC methods, whether based on attention mechanisms [5], adaptive fusion [6], or mixture of experts [7], essentially perform weighted aggregation or cross-modal interaction on all available modal features. Multi-modal data often contains substantial noise, such as irrelevant image backgrounds or generic textual descriptions. Because existing models treat all feature segments equally, they cannot dynamically focus on the most discriminative local evidence for reasoning. For example, when predicting the relation "starring in", textual features of the actor's name are far more important than descriptions of the movie's background. When predicting the entity "Eiffel Tower", the image of the tower itself is more relevant than the tourists below it. Existing methods lack a mechanism to dynamically and sparsely mask irrelevant information while focusing on key tokens. This makes models susceptible to noise interference, reducing discriminative power.

To address the above challenges, this paper proposes a new model with dynamic sparsity as its core design principle—Dynamic Sparse Selection (DSS). The core idea is to introduce dynamic sparsity as a general mechanism into both the feature fusion and contrastive learning processes. This enables precise information filtering and alignment. We propose that adding controllable sparsity can guide the model to focus dynamically on the most informative feature parts. This approach is essential for achieving precise reasoning and alignment.

Dynamic Sparse Selection (DSS)'s innovations are embodied in two core methods: Sparse-Global Fusion (S-GF) and Sparse Gated Contrastive Learning (SGCL). S-GF first tokenizes the visual and textual modalities of entities in an MMKG using pre-trained models. This captures fine-grained semantic sequences for each modality. These sequences are then fed in parallel into Sparse Prompt Attention (SPA) and Global Dense Attention (GDA). SPA, our core design, achieves

*Corresponding author.

This work was supported by the Sichuan Province Science and Technology Achievement Transfer and Transformation Demonstration Project, China (Grant No. 2024ZHCG0002)

dynamic information filtering at the fine-grained level. Concurrently, the standard GDA branch ensures the model does not lose global semantic coherence and structural dependencies. The outputs of the two branches are fused via a learnable gate. This design equips the model with both highly discriminative local feature extraction capability and robust modeling of global context.

To further enhance the alignment quality, we propose SGCL. First, this method focuses on multiple distinct perspectives of an entity. The Sparse Gated Selection Mechanism (SGSM) dynamically constructs a highly compatible anchor representation for each perspective. Subsequently, it systematically enhances the discriminative power and alignment robustness of entity representations through the Multi-view Focused Adaptive Contrastive Loss (MFAC Loss). This loss function progressively adjusts the focus on hard negative samples.

Our contributions are summarized as follows:

- We propose DSS, a novel MMKGC model based on the principle of dynamic sparsity. It is the first to systematically apply a learnable sparse mechanism to both feature selection and sample selection. This provides a unified and effective new approach for fine-grained noise filtering and precise contrastive alignment.
- We design the innovative S-GF method. The SPA branch enables key information filtering. It complements the GDA branch. This significantly improves feature discriminability while preserving structural information.
- We propose the SGCL method. A learnable sparse gate enables adaptive, dynamic selection of contrastive samples. Combined with the MFAC Loss, it markedly improves the accuracy of cross-modal alignment and the model's generalization capability.

  Due to proprietary constraints, the data and code associated with this paper cannot be made publicly available.

II. Related work

A. Multi-modal Knowledge Graph Completion

Compared to traditional KGs, Multi-modal Knowledge Graphs (MMKGs) contain richer modal information (e.g., images, text) but still suffer from incompleteness. The core challenge of Multi-modal Knowledge Graph Completion (MMKGC) lies in effectively coordinating structural, textual, visual, and other modal information to enhance entity prediction.

Early MMKGC works, such as IKRL [8] and TransAE [9], validated the effectiveness of multi-modal information by using pre-trained models to extract features from images and text, and typically combined these with structural embeddings through simple fusion operations or cross-modal loss functions. However, such coarse-grained fusion methods fail to finely distinguish the importance of information within and between different modalities, making them susceptible to interference from noise such as image backgrounds or irrelevant textual descriptions.

To optimize the fusion process, subsequent studies introduced more refined mechanisms. For example, OTKGE [10] employs optimal transport theory to minimize distributional differences between embeddings of different modalities, thereby promoting modal alignment. AdaMF [11] introduces adversarial training to balance the contributions of different modalities, enhancing robustness against modal imbalance. These methods achieve dynamic fusion at the level of the "entire image" or "entire text," improving the model's adaptability to different relational contexts. However, they still treat the feature vectors of each modality as indivisible wholes and cannot achieve differentiated attention and filtering at a finer-grained feature segment level (e.g., key object regions in images, core entity words in text).

Recently, some research has attempted finer-grained interactions. For instance, models like VISTA [12] and MKGformer [13] introduce powerful sequence modeling capabilities to MMKGC. MyGO [5] employs attention mechanisms and fine-grained encoding to tokenize modal data at a fine granularity, avoiding the information loss that arises from compressing an entire image or text into a single static embedding. Overall, such methods based on fine-grained tokens and attention share the common goal of more precisely capturing complex intra- and inter-modal dependencies. However, their attention mechanisms are typically dense, tending to assign some computational weight to all segments, even if many constitute noise. This makes it difficult for the model to actively and accurately focus on the most informative key segments and effectively suppress the influence of irrelevant segments.

In summary, existing MMKGC methods have made significant progress in the depth and dynamism of multi-modal fusion, but they still commonly face a key challenge: the lack of a dynamic, sparsity-guided discriminative information selection mechanism at the feature segment level. This prevents models from quickly focusing on the most relevant and discriminative local evidence in cross-modal signals when processing raw multi-modal data rich in noise—much like humans do. Consequently, it constrains the precision and robustness of representation learning and alignment.

III. Methodology

A. Task Formulation

An MMKG containing visual and textual modalities can be represented as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{V}, \mathcal{D})$, where $\mathcal{E}$ and $\mathcal{R}$ are the entity and relation sets. $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the set of triplets, indicating that head entity h is associated with tail entity t via relation r. $\mathcal{V}$ and $\mathcal{D}$ are the image and textual description sets corresponding to each entity $e \in \mathcal{E}$. The goal of MMKGC is to learn a scoring function $\mathcal{S}(h, r, t)$ that assigns high scores to plausible triplets and low scores to implausible ones.

B. Overall Pipelines

Our DSS model aims to enhance the ability of the MMKGC model to focus on key details in fine-grained modal information by introducing a dynamic sparse mechanism. As shown in Fig. 1, the overall workflow consists of four stages. First, in the Multi-modal Information Tokenization Stage, we use

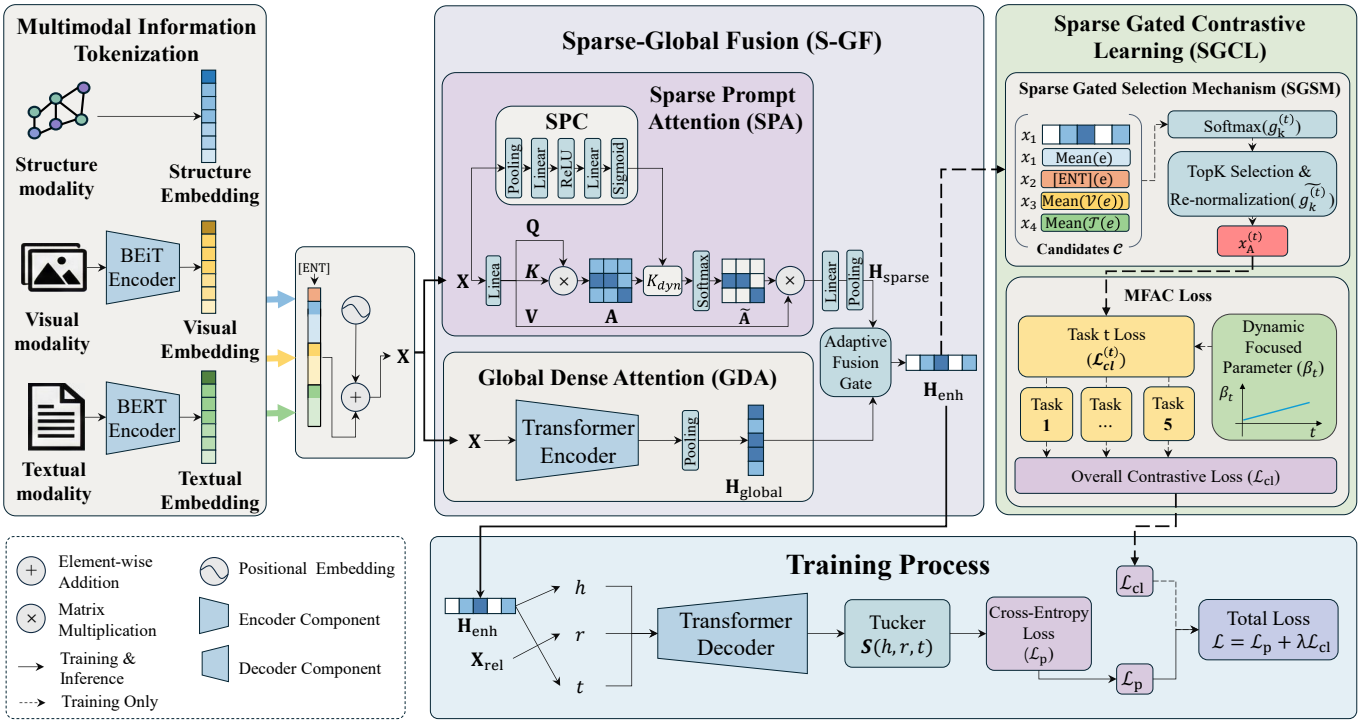

Fig. 1. The overall architecture of our proposed DSS.

pre-trained models to encode multi-modal data (text, image, and structure) into fine-grained token sequences. Second, in the S-GF Stage, the encoded sequences are fed in parallel into two branches to balance focus on local key information and modeling of complete global context. In the SPA branch, we introduce SPA to dynamically filter the most valuable modal segments (e.g., key objects or entity words) and effectively suppress noise. In the GDA branch, a standard Transformer Encoder captures the complete global context. Third, the outputs of both branches are fed into an Adaptive Fusion Gate (AFG) to generate enhanced entity representations. Finally, in the SGCL and Model Training Stage, the fused representations are optimized for cross-modal alignment via SGCL. They are also fed into a Transformer Decoder for the final triplet scoring and prediction.

The total loss function comprises two parts: a cross-entropy loss for entity prediction and a contrastive learning loss for modal alignment.

C. Multi-modal Information Tokenization

Consistent with prior MMKGC research, the multi-modal information in this paper consists of four parts: structure, text, image, and relation modalities. We use the pre-trained image encoder BEiT-V2 [14] and the pre-trained text encoder BERT [15] to extract fine-grained tokens. The formula for each modal sequence is as follows:

$$\mathbf{X}_m = \mathbf{x}_m + \mathbf{P}_{pos}, \quad m \in \{\text{str}, \text{vis}, \text{txt}, \text{rel}\} \tag{1}$$

where $\mathbf{x}_{str}$ and $\mathbf{x}_{rel}$ is the structural embedding, obtained via a learnable lookup table initialized with Xavier uniform distribution. $\mathbf{x}_{vis}$ and $\mathbf{x}_{txt}$ are the visual and textual token sequences, respectively. $\mathbf{P}_{pos}$ denotes learnable positional embeddings. The relation modality sequence $\mathbf{X}_{rel}$ is separately processed in the Transformer Decoder for triplet scoring. For an entity e, we construct a unified multi-modal sequence $\mathbf{X} = \{[\text{ENT}]; \mathbf{X}_{str}; \mathbf{X}_{vis}; \mathbf{X}_{txt}\}$, where ; denotes concatenation, and [ENT] [5] is used to capture sequence features for downstream prediction.

D. Sparse-Global Fusion Method

The Sparse-Global Fusion method is a core innovation of this study. To achieve efficient collaborative fusion of multi-modal information for enhanced entity representations, S-GF introduces a parallel architecture: a Sparse Prompt Attention branch for noise filtering, and a Global Dense Attention branch for preserving semantic and structural integrity. Finally, the outputs of the two branches are combined via an Adaptive Fusion Gate to produce the enhanced entity representation.

1) Sparse Prompt Attention: To focus on key fine-grained information in each modality, we propose Sparse Prompt Attention. The standard self-attention mechanism in Transformers computes weights for all tokens, often introducing irrelevant background noise. The core innovation of SPA is to focus only on critical cross-modal interactions, suppressing token-level noise interference in multi-modal fusion, thereby strengthening the fused entity representation.

Specifically, given the multi-modal input sequence $\mathbf{X}$, query, key, and value matrices $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are first obtained via linear projection. The global attention score matrix $\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}$ is then computed, where d is the feature dimension.

However, directly normalizing these attention scores incorporates all interactions, including those representing potential noise, affecting entity representation quality. To address this, SPA introduces a Sparse Prompt Controller (SPC). This controller adaptively determines the number of keys k_{dyn} each query should retain based on the current input. This value lies between 1 and the number of tokens, reflecting the amount of important information in the current sample.

Based on k_{dyn}, we perform dynamic sparsification on the original attention scores: only the top k_{dyn} highest attention scores for each query are retained; the rest are suppressed. This operation transforms the original score matrix $\mathbf{A}$ into a sparse attention score matrix $\widetilde{\mathbf{A}}$. Specifically, for each query row, only the largest k_{dyn} values are kept; other positions are set to negative infinity.

Thus, in the subsequent softmax calculation, the weights for suppressed positions approach zero. We obtain the sparse attention weights:

$$\text{Attn} = \text{softmax}(\widetilde{\mathbf{A}})\mathbf{V} \tag{2}$$

We adopt a multi-head strategy. The outputs of all attention heads are concatenated, linearly projected, and then passed through a global average pooling layer to obtain the enhanced entity representation $\mathbf{H}_{sparse}$.

2) Sparse Prompt Controller: The SPC module serves as the dynamic sparse controller for SPA. Its core function is to generate prompt information from the multi-modal input sequence $\mathbf{X}$, determining the number of connections k each attention head should retain, thereby facilitating the dynamic selection process for the attention matrix. SPC first compresses the sequence dimension $\mathbf{X}$ via global average pooling to obtain token-level feature statistics $\mathbf{z} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i$. $\mathbf{z}$ is then linearly projected into a hidden space, passed through a ReLU activation, and projected again linearly. A sigmoid function produces the sparse retention ratio $r \in (0, 1)$. The k_{dyn} value is finally calculated as in (3):

$$k_{dyn} = \lfloor \min(\max(r \cdot N, 1), k_{\max}) \rfloor \tag{3}$$

Here, $k_{\max}$ is the token dimension length of the multi-modal input sequence $\mathbf{X}$.

3) Adaptive Dual-Branch Integration: The sparse representation $\mathbf{H}_{\text{sparse}}$ output by the SPA branch reflects the key cross-modal interaction information filtered by the dynamic sparse attention mechanism. Simultaneously, the GDA branch processes the same multi-modal input sequence $\mathbf{X}$ via a standard Transformer encoder, followed by pooling, to obtain the global representation $\mathbf{H}_{\text{global}}$. The global branch captures complete token-to-token dependencies, ensuring no important long-range interaction information is lost.

To balance the focused features $\mathbf{H}_{\text{sparse}}$ from SPA and the holistic features $\mathbf{H}_{\text{global}}$ from the global branch, we employ an Adaptive Fusion Gate:

$$\mathbf{H}_{\text{enh}} = \beta \cdot \mathbf{H}_{\text{sparse}} + (1 - \beta) \cdot \mathbf{H}_{\text{global}} \tag{4}$$

Here, $\beta \in (0, 1)$ is a learnable parameter, constrained by a sigmoid function to the range $(0, 1)$ for training stability.

E. Sparse Gated Contrastive Learning

To enhance the discriminative ability and robustness of entity representations in MMKGC, we propose SGCL. SGCL mainly consists of two components: the SGSM and the MFAC Loss.

1) Sparse Gated Selection Mechanism: SGSM aims to dynamically construct an appropriate anchor representation for each contrastive task. Specifically, for an entity e, we build a corresponding multi-view candidate set $\mathcal{C} = \{x_1, x_2, x_3, x_4, x_5\}$. Here, $x_1 = H_{enh}(e)$ is the enhanced entity representation output by S-GF; $x_2 = [\text{ENT}]_e$ is the special token representation; $x_3 = \text{Mean}(\mathbf{X}(e))$ is the global average representation; $x_4 = \text{Mean}(\mathcal{V}(e))$ is the average visual modality representation; $x_5 = \text{Mean}(\mathcal{D}(e))$ is the average textual modality representation.

To ensure each view is fully learned and the final representation aligns with all views, we establish five contrastive learning tasks, numbered $t = \{1, 2, 3, 4, 5\}$. Each task t uses the corresponding view x_t as the positive sample. For task t, its anchor representation $x_{\text{A}}^{(t)}$ is not fixed. Instead, it is dynamically fused from the candidate set $\mathcal{C}$ via a sparse gating mechanism, using the positive sample view x_t as the query. The specific process is as follows: First, based on the cosine similarity between the positive sample view x_t and each candidate x_k ($k = 1, \ldots, 5$) in the candidate set, we compute the initial gating weights using a Softmax with a temperature parameter τ_{gate}:

$$g_k^{(t)} = \frac{\exp\left(\cos\left(x_t, x_k\right)/\tau_{\text{gate}}\right)}{\sum_{j=1}^{5}\exp\left(\cos\left(x_t, x_j\right)/\tau_{\text{gate}}\right)}, k = 1, \ldots, 5 \tag{5}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity function, and τ_{gate} is the temperature coefficient controlling the sharpness of the weight distribution.

Subsequently, to make the model focus on the most crucial candidate views, we use the Sparse Candidate Count K to determine how many high-weight candidates to retain. The weights of these selected candidates are then re-normalized to obtain the final sparse gating weights:

$$\widetilde{g_k^{(t)}} = \frac{g_k^{(t)}}{\sum i \in \text{TopK}^{(t)} g_i^{(t)}}, k \in \text{TopK}^{(t)} \tag{6}$$

where $\text{TopK}^{(t)}$ denotes the set of indices k corresponding to the top-K largest $g_k^{(t)}$ values for task t.

Finally, the anchor representation for task t is the weighted sum of the selected candidate views:

$$x_{\text{A}}^{(t)} = \sum_{k \in \text{TopK}^{(t)}} \widetilde{g_k^{(t)}} \cdot x_k \tag{7}$$

2) Multi-view Focused Adaptive Contrastive Loss: To progressively optimize the entity representation space, we propose the MFAC Loss, based on the anchor representations generated by SGSM. This loss employs a dynamically adjusted focusing parameter, enabling the model to transition from coarse-grained global alignment to finely distinguishing hard negative samples, thereby learning more discriminative representations.

For a training batch containing N entities, for each task t, we generate an anchor representation $x_{\mathrm{A}}^{(t,i)}$ for each entity $e^{(i)}$, and use the corresponding view $x_t^{(i)}$ of that entity as the positive sample.

To enhance the model's ability to distinguish hard negative samples, we introduce the Dynamic Focused Parameter β_t that dynamically increases with the training steps:

$$\beta_t = \beta_{\text{init}} + (\beta_{\text{final}} - \beta_{\text{init}}) \cdot \min(t/T, 1) \tag{8}$$

where t is the current training epoch, T is the preset total training epochs, β_{init} is the initial β_t value, and β_{final} is the final β_t value. In the early stages of training, β_t is small. The loss function approximates the standard contrastive loss, with the primary goal of pulling positive pairs closer, encouraging the model to learn global, coarse-grained alignment. As training progresses, β_t increases. The loss function then imposes stronger penalties on negative samples with high similarity to the anchor, forcing the model to focus on distinguishing these hardest samples.

For task t, its contrastive loss function is defined as follows:

$$\mathcal{L}_{\text{cl}}^{(t)} = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{\exp\left(\cos\left(x_{\mathrm{A}}^{(t,i)}, x_t^{(i)}\right)/\tau_{\text{cl}}\right)}{\exp\left(\cos\left(x_{\mathrm{A}}^{(t,i)}, x_t^{(i)}\right)/\tau_{\text{cl}}\right) + \sum_{j\neq i}\exp\left(\beta_t \cdot \cos\left(x_{\mathrm{A}}^{(t,i)}, x_t^{(j)}\right)/\tau_{\text{cl}}\right)}\right] \tag{9}$$

where τ_{cl} is the temperature parameter.

Finally, the overall contrastive loss is the average of the five task losses:

$$\mathcal{L}_{\text{cl}} = \frac{1}{5}\sum_{t=1}^{5}\mathcal{L}_{\text{cl}}^{(t)} \tag{10}$$

F. Training Process

This method employs a Transformer decoder-based triple scoring module. We extract the vectors of entities and relations from $\mathbf{H}_{\text{enh}}$ and x_{rel} respectively, transforming the candidate triple (h, r, t) into a vector sequence $[e_h; e_r; e_t]$ for subsequent processing. This sequence is then fed into a multi-layer Transformer decoder for encoding, which outputs the corresponding contextual vector sequence. Subsequently, we introduce Tucker [16] as our scoring function $\mathcal{S}(h, r, t) = \mathcal{W} \times_1 \widetilde{h} \times_2 \widetilde{r} \times_3 t$, where $\times_i$ denotes the vector product along the i-th mode, and $\mathcal{W}$ is a learnable tensor during training. We use this scoring function to produce a scalar score measuring the plausibility of the triple. Finally, the obtained scalar scores are used to compute the standard Cross-Entropy Loss $\mathcal{L}_p$:

TABLE I
THE STATISTICAL INFORMATION OF THE DB15K AND MKG-W/Y DATASETS.

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	#Vis	#Text	Train	#Valid	#Test
DB15K	12842	279	12818	12842	79222	9902	9904
MKG-W	15000	169	14463	14123	34196	4276	4274
MKG-Y	15000	28	14244	14305	21310	2665	2663

$$\mathcal{L}_p = -\sum_{(h,r,t)\in\mathcal{T}} log\frac{\exp\left(\mathcal{S}\left(h,r,t\right)\right)}{\sum_{t'\in\mathcal{E}}\exp\left(\mathcal{S}\left(h,r,t'\right)\right)} \tag{11}$$

Here, we treat t as the ground truth label of the entity set $\mathcal{E}$, which is consistent with the handling of h prediction.

Finally, the overall objective for the MMKGC task is $\mathcal{L} = \mathcal{L}_p + \lambda\mathcal{L}_{cl}$, where λ is the Loss Weight for $\mathcal{L}_{cl}$.

IV. EXPERIMENTS

A. Experimental Setup

1) Dataset: This paper evaluates the model using three public MMKGC benchmarks: DB15K [17] and MKG-W/Y [18]. The original data for each modality is obtained from official sources. The statistical information of the three datasets is shown in Table I.

2) Evaluation Protocol: We perform the link prediction task on the dataset. This task requires predicting the missing entity in the given queries $(h, r, ?)$ or $(?, r, t)$ from $\mathcal{T}_{\text{test}}$. We apply the filtered setting to the datasets to filter out already existing facts. This paper employs the same metrics as prior research for result evaluation: Mean Reciprocal Rank (MRR) and Hit@N (N=1, 3, 10). These metrics are calculated as follows: $\text{MRR} = \frac{1}{|\mathcal{T}_{\text{test}}|}\sum_{i=1}^{|\mathcal{T}_{\text{test}}|}\left(\frac{1}{r_{h,i}} + \frac{1}{r_{t,i}}\right)$, $\text{Hit@}N = \frac{1}{|\mathcal{T}_{\text{test}}|}\sum_{i=1}^{|\mathcal{T}_{\text{test}}|}\left[\mathbb{1}(r_{h,i} \leq K) + \mathbb{1}(r_{t,i} \leq K)\right]$, where $r_{h,i}$ and $r_{t,i}$ are the ranks for head and tail predictions, respectively.

3) Baseline Models: We select 17 representative MMKGC methods as baselines. (1) Uni-modal KGC methods: TransE [19], ComplEx [20], RotatE [21], QuatE [22], DualE [23]. These models only incorporate the structural information of the knowledge graph. (2) Multi-modal KGC models: IKRL [8], TransAE [9], VBKGC [24], OTKGE [10], MoSE [25], MMRNS [26], QEB [27], VISTA [12], IMF [28], AdaMF [11], MyGO [5], K-ON [29]. These methods leverage both the structural and multi-modal information in KGs.

4) Implementation Details: In our experiments, we implement DSS using PyTorch and utilize an RTX 4090 GPU. For modal tokenization, we adopt BEiT-V2 [14] and BERT [15] as the visual and textual tokenizers. The codebook size for BEiT-V2 is 8192, and the vocabulary size for the BERT tokenizer is 32,000. During training, we set the number of epochs to 2000, the batch size to 2048, and the embedding dimension to 256. The maximum number of tokens for text and vision is 12. The weight parameter λ is tuned within the range $\{1\text{e}{-3}, 1\text{e}{-4}, 1\text{e}{-5}, 1\text{e}{-6}\}$. The model is optimized using the Adam optimizer [30].

B. Main Results

DSS outperforms all 17 baselines across multiple metrics, with performance improvements ranging from 0.6% to 4.6%, highlighting its ability to fuse fine-grained multi-modal information effectively. As shown in Table II, these results demonstrate that the sparse mechanism enhances alignment and boosts prediction accuracy. Furthermore, we observe that the DSS model shows larger relative improvements on the Hit@1 and MRR metrics across the three datasets, with an average increase of 2.61%. This indicates the model's excellent capability for high-precision prediction.

C. Ablation Study

1) Model Design: To verify the effectiveness of each module in DSS, we conduct further ablation studies from the perspectives of Model Design.

The ablation studies confirm that both the S-GF and SGCL modules are critical to the overall performance of DSS, with the SGCL contributing more significantly. As shown in Table III, removing the S-GF leads to a drop of 1.52% in MRR and 2.78% in Hit@1, whereas removing the SGCL causes a more substantial decline of 3.88% in MRR and 5.40% in Hit@1. This result strongly demonstrates the substantial contribution of S-GF and SGCL. The two methods enables precise multi-modal feature alignment and improves model prediction accuracy.

Ablating Dynamic Sparsity: Dynamic sparsity is necessary for adaptive fusion. We ablate the dynamic sparse mechanism (w/o Dynamic Sparsity) by replacing the SPA in SPC and the SGSM in SGCL with uniform aggregation. In this setting, MRR and Hit@1 drop by an average of 2.10%, indicating that static, non-adaptive fusion cannot replace the dynamic sparse filtering mechanism.

Ablating Sub-modules in SGCL: The two sub-modules in SGCL work synergistically to enhance the overall performance. Removing either the SGSM or the MFAC Loss leads to an average drop of 3.45% in MRR and 4.37% in Hit@1. This shows that the effectiveness of SGCL relies on its complete dynamic architecture: the SGSM performs hard selection at the sample level, while the MFAC Loss achieves progressive focusing at the loss function level. Their synergy is essential for optimal alignment and accurate prediction.

D. Impact of Dual Branch Attention

As shown in Table IV, both SPA and GDA are crucial for the MMKGC task, as confirmed by our experiments. Optimal performance is achieved only when both attention mechanisms are used concurrently. Rows 1, 2, and 3 represent three different ablation scenarios: removing both SPA and GDA from the S-GF module, removing only SPA, and removing only GDA, respectively. All three scenarios resulted in varying degrees of performance degradation. This result demonstrates the complementary roles and individual necessity of the two attention mechanisms.

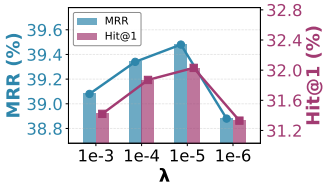

Fig. 2. Different values of λ.

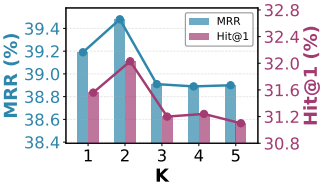

Fig. 3. Different values of K.

E. Hyper-parameter Sensitivity Experiments

To further investigate the sensitivity of important parameters in the contrastive learning component, we examine the impact of the SGCL Loss Weight λ, the Sparse Candidate Count K, and the Dynamic Focused Parameter β_t on DSS, using the DB15K dataset.

1) Loss Weight λ: The experimental results are shown in Fig. 2. The optimal MMKGC performance is achieved at $\lambda = 1e-5$, as performance exhibits a trend of first increasing and then decreasing when λ is adjusted within $\{1e-3, 1e-4, 1e-5, 1e-6\}$.

2) Sparse Candidate Count K: As shown in Fig. 3, The best MRR and Hit@1 performance is achieved when K is 2. This result strongly supports the necessity of the sparse mechanism in SGCL. Specifically, when $K = 1$, performance declines because the anchor representation is generated solely from a single view, failing to effectively integrate key information from other perspectives. Conversely, a larger K leads to a decrease in metrics because it introduces more candidates with lower information content. Blindly increasing K for synthesizing negative samples does not yield benefits and can even cause performance degradation. The experimental results powerfully demonstrate that introducing sparsity into contrastive learning plays a crucial role in improving model performance.

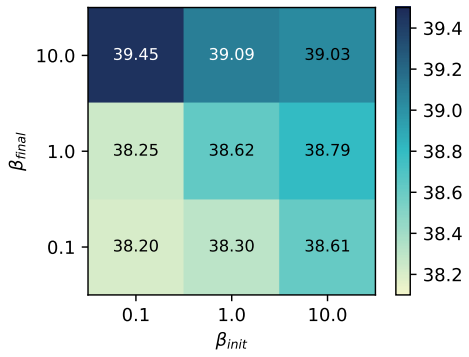

Fig. 4. The MRR performance of the DSS model on the DB15K dataset with different β_{init} and β_{final}.

3) Dynamic Focused Parameter β_t: We analyze the impact of different β_{init} and β_{final} settings on DSS performance using the DB15K dataset, where β_{init} and β_{final} are varied over a grid of values $\{0.1, 1.0, 10.0\}$. The experimental results are shown in Fig 4.

TABLE II
THE MAIN MMKGC RESULTS. THE BEST RESULTS ARE MARKED AS **BOLD** AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Model	DB15K				MKG-W				MKG-Y			
	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37
ComplEx	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73	28.71	22.26	32.12	40.93
RotatE	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30
QuatE	34.18	25.42	38.91	51.30	34.50	28.94	36.71	46.64	36.01	30.53	38.84	43.68
DualE	35.85	29.31	38.52	51.28	33.94	27.55	36.56	46.09	34.95	29.77	38.44	43.12
IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07	33.22	30.37	34.28	28.26
TransAE	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03
VBKGC	30.61	19.75	37.18	49.44	30.61	24.91	33.01	40.88	37.04	33.76	38.75	42.30
OTKGE	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88	35.51	31.97	37.18	41.38
MoSE	28.38	21.56	30.91	41.67	33.34	27.78	33.94	41.06	36.28	33.64	37.47	40.81
MMRNS	32.68	23.01	37.86	51.01	35.03	28.59	37.49	47.47	35.93	30.53	39.07	45.47
QEB	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32	34.37	29.49	36.95	42.32
VISTA	30.42	22.39	33.56	45.94	32.91	26.12	35.38	45.61	30.45	24.87	32.39	41.53
IMF	32.25	24.20	36.00	48.19	34.50	28.77	36.62	45.44	35.79	32.95	37.14	40.63
AdaMF	32.51	21.31	39.67	51.68	34.27	27.21	37.86	47.21	38.06	33.49	**40.44**	**45.48**
MyGO	37.72	30.08	41.26	52.21	36.10	29.78	**38.54**	47.75	38.44	35.01	39.84	44.19
K-ON	36.24	28.13	40.49	51.26	35.83	29.41	37.32	47.16	35.83	32.56	37.34	42.45
Ours	**39.48**	**32.03**	**43.06**	**53.80**	**36.29**	**30.60**	37.88	**47.79**	**38.54**	**35.35**	39.92	43.86

TABLE III
THE ABLATION STUDY RESULTS ON DB15K.

Setting	MRR	Hit@1	Hit@3	Hit@10
w/o S-GF	38.88	31.14	42.51	53.76
w/o SGCL	37.95	30.30	41.39	52.76
w/o Dynamic Sparsity	38.77	31.26	42.42	53.20
w/o SGSM IN SGCL	38.09	30.60	41.48	52.45
w/o MFAC Loss IN SGCL	38.15	30.66	41.56	52.61
Full Model	39.48	32.03	43.06	53.80

TABLE IV
ABLATION STUDY OF DUAL BRANCH ATTENTION ON DB15K.

SPA	GDA	DB15K			
		MRR	Hit@1	Hit@3	Hit@10
✗	✗	38.88	31.14	42.51	53.76
✗	✓	39.11	31.31	42.86	54.22
✓	✗	37.86	30.38	41.40	51.95
✓	✓	39.48	32.03	43.06	53.80

Firstly, employing a monotonically increasing β_t schedule is crucial for achieving optimal performance. The model attains the highest MRR when setting $\beta_{\text{init}} = 0.1$ and $\beta_{\text{final}} = 10.0$. This optimal configuration allows β_t to linearly increase from a low to a high intensity. It follows the progressive focusing strategy designed for the MFAC Loss. Specifically, it initially focuses on pulling positive sample pairs closer for global alignment in the early training stages. Subsequently, it gradually strengthens the discrimination of hard negative samples.

Secondly, the final focusing intensity β_{final} is a key factor determining the model's discriminative ability. Experiments show that, given the same β_{init}, increasing β_{final} consistently yields performance gains. For example, with $\beta_{\text{init}} = 0.1$, increasing β_{final} from 0.1 to 10.0 improves MRR by 3.27%. This confirms that applying stronger contrastive penalties in the later training stages is essential for learning a highly discriminative representation space.

In summary, the experiments on the dynamic focused parameter β_t validate the important role of the MFAC Loss. Its design, through an adaptive contrastive learning process that transitions from global alignment to fine-grained discrimination, works synergistically with the SGSM. This joint mechanism ensures the model robustly learns discriminative and resilient multi-modal entity representations.

V. CONCLUSION

In this paper, we propose DSS, a Multi-modal Knowledge Graph Completion model based on Dynamic Sparse Selection. Specifically, we design the S-GF method. Within it, SPA can dynamically filter the most discriminative feature segments across modalities, effectively suppressing image background noise and textual redundancy. Simultaneously, Global Dense Attention preserves the complete global context, ensuring semantic coherence and structural integrity. The two are fused via an adaptive gating mechanism, achieving synergy between local focus and global modeling. Furthermore, we

propose SGCL. It uses a learnable sparse gate to dynamically select the most informative multi-modal view features. Combined with the MFAC Loss for progressive optimization, it significantly enhances the discriminative power of entity representations and the robustness of cross-modal alignment. Extensive experiments on three benchmark datasets, DB15K, MKG-W, and MKG-Y, demonstrate that DSS outperforms existing methods across multiple evaluation metrics. This validates the effectiveness and generalization capability of the proposed dynamic sparse filtering mechanism and sparse gated contrastive strategy.

REFERENCES

[1] Z. Chen, Y. Zhang, Y. Fang, Y. Geng, L. Guo, X. Chen, Q. Li, W. Zhang, J. Chen, Y. Zhu, J. Li, X. Liu, J. Z. Pan, N. Zhang, and H. zeng Chen, "Knowledge graphs meet multi-modal learning: A comprehensive survey," *ArXiv*, vol. abs/2402.05391, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267547866

[2] K. Y. Liang, L. Meng, M. Liu, Y. Liu, W. Tu, S. Wang, S. Zhou, X. Liu, and F. Sun, "A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 9456–9478, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:257220329

[3] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng, "Multi-modal knowledge graphs for recommender systems," *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:224281034

[4] Y. Zhu, H. Tou, W. Zhang, G. Ye, H. Chen, N. Zhang, and H. Chen, "Knowledge perceived multi-modal pretraining in e-commerce," *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237386166

[5] Y. Zhang, Z. Chen, L. Guo, Y. Xu, B. Hu, Z. Liu, H. zeng Chen, and W. Zhang, "Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation," in *AAAI Conference on Artificial Intelligence*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269148843

[6] Y. Zhang, Z. Chen, L. Guo, Y. Xu, B. Hu, Z. Liu, W. Zhang, and H. zeng Chen, "Native: Multi-modal knowledge graph completion in the wild," *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270711106

[7] L. Li, "Complementarity-driven representation learning for multi-modal knowledge graph completion," *ArXiv*, vol. abs/2507.20620, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:280323661

[8] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *International Joint Conference on Artificial Intelligence*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:9909815

[9] Z. Wang, L. Li, Q. Li, and D. D. Zeng, "Multimodal data enhanced representation learning for knowledge graphs," *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:203605587

[10] Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang, "Otkge: Multi-modal knowledge graph embeddings via optimal transport," in *Neural Information Processing Systems*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:258509157

[11] Y. Zhang, Z. Chen, L. Liang, H. zeng Chen, and W. Zhang, "Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion," *ArXiv*, vol. abs/2402.15444, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267898006

[12] J. Lee, C. Chung, H. Lee, S. Jo, and J. J. Whang, "Vista: Visual-textual knowledge graph representation learning," in *Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266166905

[13] X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, and H. Chen, "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248524814

[14] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *ArXiv*, vol. abs/2106.08254, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235436185

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[16] I. Balazevic, C. Allen, and T. M. Hospedales, "Tucker: Tensor factorization for knowledge graph completion," *ArXiv*, vol. abs/1901.09590, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:59316623

[17] Y. Liu, H. Li, A. García-Durán, M. Niepert, D. Oñoro-Rubio, and D. S. Rosenblum, "Mmkg: Multi-modal knowledge graphs," *ArXiv*, vol. abs/1903.05485, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:76663467

[18] D. Xu, T. Xu, S. Wu, J. Zhou, and E. Chen, "Relation-enhanced negative sampling for multimodal knowledge graph completion," *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252783084

[19] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Neural Information Processing Systems*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:14941970

[20] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," *ArXiv*, vol. abs/1606.06357, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:15150247

[21] Z. Sun, Z. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," *ArXiv*, vol. abs/1902.10197, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:67855617

[22] S. Zhang, Y. Tay, L. Yao, and Q. Liu, "Quaternion knowledge graph embeddings," in *Neural Information Processing Systems*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202763266

[23] Z. Cao, Q. Xu, Z. Yang, X. Cao, and Q. Huang, "Dual quaternion knowledge graph embeddings," in *AAAI Conference on Artificial Intelligence*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235306387

[24] Y. Zhang and W. Zhang, "Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling," *ArXiv*, vol. abs/2209.07084, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252280329

[25] Y. Zhao, X. Cai, Y. Wu, H. Zhang, Y. Zhang, G. Zhao, and N. Jiang, "Mose: Modality split and ensemble for multimodal knowledge graph completion," in *Conference on Empirical Methods in Natural Language Processing*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252918783

[26] D. Xu, T. Xu, S. Wu, J. Zhou, and E. Chen, "Relation-enhanced negative sampling for multimodal knowledge graph completion," *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252783084

[27] X. Wang, B. Meng, H. Chen, Y. Meng, K. Lv, and W. Zhu, "Tiva-kg: A multimodal knowledge graph with text, image, video and audio," *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264492337

[28] X. Li, X. Zhao, J. Xu, Y. Zhang, and C. Xing, "Imf: Interactive multimodal fusion model for link prediction," *Proceedings of the ACM Web Conference 2023*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257631615

[29] L. Guo, Y. Zhang, Z. Bo, Z. Chen, M. Sun, Z. Zhang, W. Zhang, and H. zeng Chen, "K-on: Stacking knowledge on the head layer of large language model," *ArXiv*, vol. abs/2502.06257, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276249512

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106