DSF^2: DUAL-SPACE GUIDED DISCRIMINATIVE HIGH AND LOW FREQUENCY FUSION NETWORK FOR MULTIMODAL KNOWLEDGE GRAPH COMPLETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal knowledge graph completion (MMKGC) aims to complete missing facts in knowledge graphs by jointly leveraging multimodal information associated with entities. Existing methods primarily fuse multimodal entity embeddings in the spatial domain, yet lack explicit modeling of frequency characteristics at different scales of variation. To capture these multiscale characteristics, shifting to the frequency domain provides a new perspective by employing spectral decomposition and filtering to facilitate multimodal fusion, but such methods neglect the pervasive coupling of spectral signals at the boundaries and the distinct characteristics of different frequencies, thereby hampering the precise characterization of multimodal features and limiting the exploitation of multimodal spectral information. To address these limitations, we propose DSF^2, Dual-Space Guided Discriminative High and Low Frequency Fusion Network for MMKGC, an innovative framework that leverages different spatial geometrical properties to disentangle the pervasive inter-frequency coupling and applies distinct modeling strategies tailored to band-specific characteristics. Concretely, we design Dual-Space Frequency Disambiguation (DSFD), which better models the coupling of spectral signals by leveraging the geometric differences of corresponding spectral embeddings across these spaces to derive more discriminative features. We further design distinct processing strategies for differentiated frequencies, including a low-frequency Shared-memory Interaction Strategy (LSIS) and a High-frequency Relation-aware Gating Strategy (HRGS). Specifically, LSIS captures stable global low-frequency semantics through shared-memory interaction to promote cross-modal semantic communication and enhance multimodal consistency. HRGS focuses on high-frequency semantics characterized by local variations, employing a relation-aware gating mechanism to emphasize local information discrepancies and strengthen relation-specific high-frequency representations. This distinct strategy effectively captures differentiated frequency characteristics, thereby fully exploiting the complementary low- and high-frequency semantics. Extensive experiments on multiple benchmark datasets consistently validate the effectiveness and robustness of DSF^2. Our code is available at https://anonymous.4open.science/r/DSF2-E792/.

1 INTRODUCTION

Multimodal knowledge graphs (MMKGs) (Chen et al., 2024) extend conventional knowledge graphs (KGs) (Liang et al., 2024) by associating structured relational facts with heterogeneous modality information, such as entity images and textual descriptions. By jointly organizing symbolic knowledge and unstructured multimodal content, MMKGs provide a richer representation of real-world entities and relations. Learning entity representations in MMKGs requires jointly modeling structural dependencies and heterogeneous visual and textual information. To predict missing links, multimodal knowledge graph completion (MMKGC) methods can employ adaptive fusion (Zhang et al., 2024b), through which visual and textual features are integrated with structural embeddings to exploit complementary cues across modalities.

Existing MMKGC methods primarily operate in the spatial domain and can be categorized by their prediction mechanisms. One line focuses on discriminative completion (Zhang et al., 2024a; 2025;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Jian et al., 2025; Chen et al., 2025; Liu et al., 2026b). Another line focuses on generative and hybrid completion (Wang et al., 2026; Xiao & Zhang, 2026; Cui et al., 2026). Although these methods improve the utilization of multimodal information, existing spatial-domain approaches overlook the smooth semantics encoded at low frequencies and the local-variation semantics captured by high-frequency components, thereby limiting their ability to exploit latent frequency information.

Recent studies have introduced frequency-domain modeling to capture the spectral characteristics of entity features. One line focuses on frequency-domain methods for MMKGC (Xu et al., 2025; Wang et al., 2025a; Nguyen et al., 2026). Another line focuses on frequency-domain methods for structural KGC (Wei & Mei, 2023; Vu et al., 2023; Wang et al., 2025b). However, existing frequency-domain methods still face two limitations. First, existing frequency-domain methods generally lack explicit modeling of representations with pervasive coupling of frequency signals near the low and high-frequency boundary. Second, low-frequency usually capture smooth and stable semantics, while high-frequency describe local variations and modality-specific information. As illustrated in Figure 1, existing methods adopt uniform processing strategies, without designing distinct processing strategies according to the semantic differences between low- and high-frequency.

To address these limitations, we propose DSF^2, Dual-Space Guided Discriminative High and Low Frequency Fusion Network for MMKGC, which leverages proximity differences between corresponding frequency polarities across distinct spaces to disentangle frequency signals and employs differentiated modeling paradigms tailored to the heterogeneous characteristics of distinct frequency. DSF^2 includes three important components: Dual-Space Frequency Disambiguation (DSFD), low-frequency shared-memory interaction strategy (LSIS) and high-frequency relation-aware gating strategy (HRGS). Specifically, in DSFD, ambiguous entity-modality instances are mapped into Poincaré space, where we use hyperbolic distances to clear low- and high-frequency references to constrain their embeddings, thereby shaping their proximity to these references and supporting the disambiguation of coupled frequency components. Building upon this, DSFD performs cross-space contrastive learning, exploiting proximity differences across distinct spaces to further separate frequency components and obtain more discriminative frequency. To fully exploit the smooth and consistent semantics at low frequencies, LSIS introduces modality-specific communicators and a shared memory bank through which structural, visual, and textual low-frequency representations are written, read, and updated, facilitating cross-modal sharing, interaction, and accumulation of stable semantics. Relation-discriminative local variations are preserved through HRGS, which performs relation-aware multi-scale filtering within each modality followed by cross-modal filtering, such that frequency cues most informative for the current relation are retained while less relevant local variations are suppressed. Shared-memory interaction is further regulated by relation-aware high-frequency cues so that stable semantics can be shared while modality specificity is preserved and low-frequency homogenization is avoided. Our contributions are as follows:

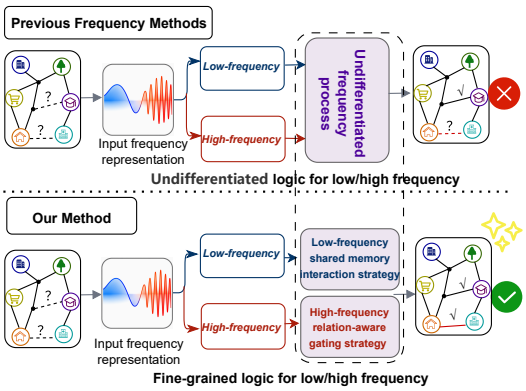


Figure 1: Comparison between existing frequency-domain methods and DSF^2.

- DSFD exploits geometric differences across distinct spaces to guide cross-space spectral discrimination, thereby overcoming the limitations of the uniform Euclidean metric and enabling the disambiguation of frequency-signal coupling.
- LSIS is proposed to enhance cross-modal interaction and accumulation among structural, visual, and textual information through shared-memory writing, reading, and updating.
- To exploit high-frequency features, HRGS performs relation-aware multi-scale filtering within each modality, followed by cross-modal filtering. HRGS captures relation-specific high-frequency semantics and guides shared-memory interaction in LSIS.
- Extensive evaluations are conducted on multiple benchmarks. DSF^2 consistently outperforms recent spatial- and frequency-domain baselines demonstrating its effectiveness.

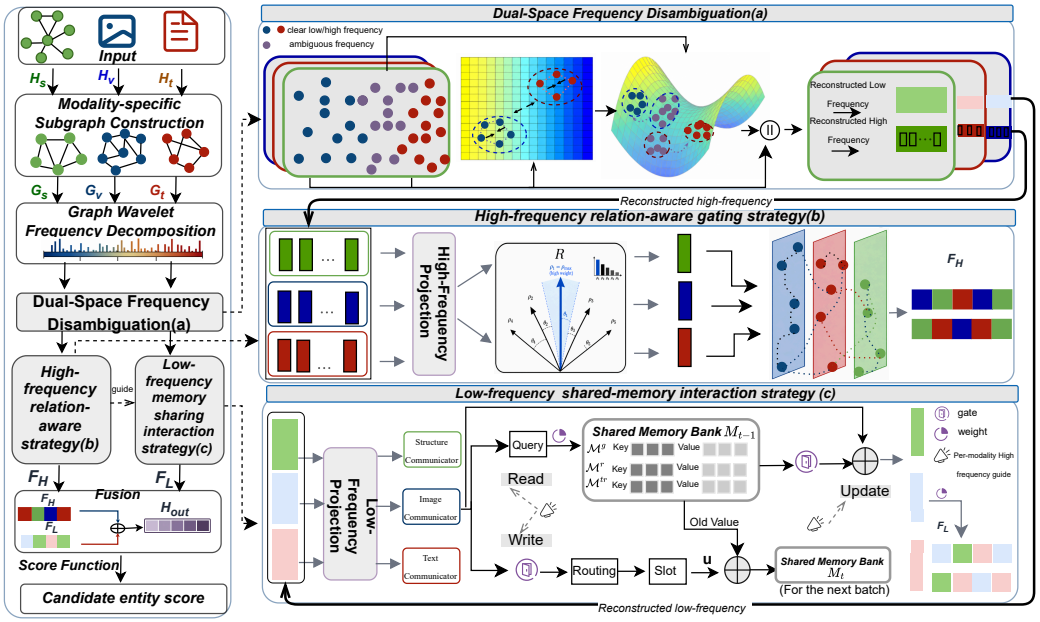


Figure 2: Overall architecture of DSF^2

2 Related Work

Recent MMKGC studies have progressively explored more adaptive and fine-grained multimodal interaction mechanisms. AdaMF-MAT (Zhang et al., 2024b) addresses modality imbalance through adaptive fusion and adversarial training. MCKGC (Gao et al., 2025) models multimodal information in mixed-curvature spaces with progressive fusion. VL-KGE (Efthymiou et al., 2026) leverages vision-language models for cross-modal alignment and integrates the aligned representations with structured relational modeling. Further details of existing MMKGC methods are provided in Appendix A.1. Though the aforementioned methods mainly operate in the spatial domain, they fail to explicitly characterize frequency-specific semantics manifested at different scales of variation, thereby overlooking the smooth and consistent semantics encoded in low frequency as well as the local-variation semantics captured by high frequency.

The introduction of frequency-domain modeling offers a new perspective for MMKGC, in which multimodal representations can be characterized according to their variation patterns across different frequency scales. By decomposing entity representations into components with distinct frequency characteristics, heterogeneous semantic cues that are difficult to disentangle in the spatial domain can be more explicitly exposed. WFF (Xu et al., 2025) applies discrete wavelet transform for frequency decomposition, filtering, and cross-modal fusion. MLFormer (Wang et al., 2025a) combines filter gating and Fourier transform for multimodal representation learning. MMQuat (Nguyen et al., 2026) uses quaternion Fourier transform filtering to reduce redundant modality information. Please refer to Appendix A.2 for more detailed studies. Despite the promise of frequency-domain modeling, existing methods lack explicit modeling of the pervasive coupling of frequency signals and distinct processing of different frequency semantics. Our method instead separates clear and ambiguous frequencies, with dedicated interactions for low-frequency shared semantics and high-frequency relation-aware semantics.

3 Method

As illustrated in Figure 2, there are three key components in DSF^2: DSFD explicitly models clear and boundary-ambiguous entity–modality instances in Euclidean and Poincaré spaces, respectively, LSIS performs shared-memory interaction to accumulate stable low-frequency semantics across modalities, whereas HRGS performs relation-aware high-frequency selection and provides discriminative guidance for low-frequency communication.

3.1 PRELIMINARIES

An MMKG is represented as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{M})$, where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ are the entity set, relation set, and observed triples, respectively. $\mathcal{M}$ comprises structural, visual, and textual modalities, with $\mathbf{X}^m(e)$ denoting the information of entity $e \in \mathcal{E}$ in modality $m \in \mathcal{M}$. MMKGC aims to learn a scoring function $S(h, r, t) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$ to assess triple plausibility and rank candidate entities for queries $(?, r, t)$ or $(h, r, ?)$.

3.2 MULTIMODAL FEATURE ENCODING AND GRAPH WAVELET DECOMPOSITION

For each modality $m \in \{s, v, t\}$ (structural, visual, and textual), entity features are projected and layer-normalized to obtain the graph signal $\mathbf{H}_m$. Structural embeddings are learned from observed KG triples, while visual and textual features are extracted using pretrained encoders. We construct modality-specific graphs $\mathcal{G}_m = (\mathcal{E}, \mathbf{A}_m)$ over the shared entity set $\mathcal{E}$. The adjacency matrix $\mathbf{A}_s$ is derived from observed KG triples, whereas $\mathbf{A}_v$ and $\mathbf{A}_t$ are constructed using cosine similarity of the corresponding modality features. To capture frequency-specific semantics, we apply spectral graph wavelet decomposition (Hammond et al., 2011). Each normalized graph Laplacian is $\mathbf{L}_m = \mathbf{I} - \mathbf{D}_m^{-1/2}\mathbf{A}_m\mathbf{D}_m^{-1/2} = \mathbf{U}_m\mathbf{\Lambda}_m\mathbf{U}_m^\top$, where $\mathbf{D}_m$ is the degree matrix, $\mathbf{U}_m$ contains the eigenvectors, and $\mathbf{\Lambda}_m$ contains the eigenvalues. For the normalized Laplacian, $\lambda_i \in [0, 2]$, where smaller eigenvalues (close to 0) correspond to smooth low-frequency patterns and larger ones (close to 2) represent high-frequency variations. Applying the spectral kernels yields the low-frequency and multi-scale high-frequency representations:

$$\mathbf{C}_m^L = \mathbf{U}_m h(\mathbf{\Lambda}_m)\mathbf{U}_m^\top \mathbf{H}_m, \qquad \mathbf{C}_m^{H,j} = \mathbf{U}_m g(s_j\mathbf{\Lambda}_m)\mathbf{U}_m^\top \mathbf{H}_m, \quad j = 1, \ldots, J. \tag{1}$$

Here, $h(\cdot)$ is the low-frequency kernel, $g(s_j\cdot)$ is the high-frequency kernel at scale s_j, and J is the number of high-frequency scales. Implementation details are provided in Appendix B.

3.3 DUAL-SPACE FREQUENCY DISAMBIGUATION

When low- and high-frequency wavelet kernels overlap in their transition region, spectral coefficients within the overlapping band can receive non-negligible responses from both kernels, leading to spectral boundary coupling. Motivated by this overlap, DSFD adopts an energy-based heuristic to identify entity-modality instances with ambiguous frequency dominance for cross-space refinement. The learned ambiguity-aware geometry is subsequently mapped back to the tangent space and combined with the original wavelet components to reconstruct the corresponding low- and multi-scale high-frequency representations. Specifically, we compute the low- and high-frequency response energies as follows:

$$E_{k,m}^L = \|\mathbf{c}_{k,m}^L\|_2^2, \quad E_{k,m}^H = \textstyle\sum_{j=1}^J \omega_j \|\mathbf{c}_{k,m}^{H,j}\|_2^2, \quad \textstyle\sum_{j=1}^J \omega_j = 1. \tag{2}$$

Here, $\mathbf{c}_{k,m}^L$ and $\mathbf{c}_{k,m}^{H,j}$ denote the low- and high-frequency representations of entity e_k in modality m, respectively, and ω_j is the weight of the j-th high-frequency scale. Based on the energy difference, we define the frequency polarity as:

$$q_{k,m} = \frac{E_{k,m}^H - E_{k,m}^L}{E_{k,m}^H + E_{k,m}^L + \epsilon}, \qquad \begin{aligned} \mathcal{B}_L &= \{(k,m) \mid q_{k,m} < -\delta\} \\ \mathcal{B}_H &= \{(k,m) \mid q_{k,m} > \delta\} \\ \mathcal{B}_A &= \{(k,m) \mid |q_{k,m}| \le \delta\} \end{aligned} \tag{3}$$

Here, $q_{k,m}$ measures the relative dominance of high- and low-frequency responses.We use $|q_{k,m}| \le \delta$ as a heuristic criterion for selecting instances with ambiguous frequency dominance, rather than as a direct measurement of spectral boundary coupling. $\epsilon > 0$ is a small constant for numerical stability. Accordingly, $\mathcal{B}_L$ and $\mathcal{B}_H$ contain low- and high-frequency-dominant instances, respectively, while $\mathcal{B}_A$ contains the selected ambiguous instances. We define $\mathcal{B}_C = \mathcal{B}_L \cup \mathcal{B}_H$. Further analysis is provided in Appendix C. To facilitate dual-space frequency disambiguation, we construct a unified frequency representation $\mathbf{f}_{k,m} = [\mathbf{c}_{k,m}^L \| \mathbf{c}_{k,m}^{H,1} \| \ldots \| \mathbf{c}_{k,m}^{H,J} \| q_{k,m}]$, where $\mathbf{f}_{k,m}$ contains the complete frequency information of entity e_k in modality m, including low-frequency semantics, multi-scale high-frequency variations, and frequency polarity. According to the energy-based grouping criterion, clear-frequency instances $(k, m) \in \mathcal{B}_C$ are mapped into Euclidean space to obtain $\mathbf{z}_{k,m}^E$, while selected ambiguous instances $(k, m) \in \mathcal{B}_A$ are mapped into Poincaré space to obtain $\mathbf{z}_{k,m}^P$.

To enable cross-space comparison, the refined clear-frequency representation is mapped into Poincaré space as $\widetilde{\mathbf{z}}^P_{k,m} = \exp_0^c\left(\mathrm{sg}(\mathbf{z}^E_{k,m})\right)$, where stop-gradient is applied to preserve the optimized Euclidean representation. For boundary-ambiguous instances, Poincaré-space frequency disambiguation is performed using $\mathbf{z}^P_{k,m}$ as the reference. The dual-space frequency disambiguation objective is formulated as

$$\mathcal{L}_{\mathrm{DSFD}} = -\frac{1}{|\mathcal{B}_C|}\sum_{(k,m)\in\mathcal{B}_C}\log\frac{\sum_{(k',m')\in\mathcal{P}^{E,+}_{k,m}} s_E\left(\mathbf{z}^E_{k,m},\mathbf{z}^E_{k',m'}\right)}{\sum_{(k',m')\in\mathcal{P}^{E,+}_{k,m}\cup\mathcal{P}^{E,-}_{k,m}} s_E\left(\mathbf{z}^E_{k,m},\mathbf{z}^E_{k',m'}\right)} - \frac{1}{|\mathcal{B}_A|}\sum_{(k,m)\in\mathcal{B}_A}\log\frac{\sum_{(k',m')\in\mathcal{P}^{P,+}_{k,m}} s_P\left(\mathbf{z}^P_{k,m},\widetilde{\mathbf{z}}^P_{k',m'}\right)}{\sum_{(k',m')\in\mathcal{P}^{P,+}_{k,m}\cup\mathcal{P}^{P,-}_{k,m}} s_P\left(\mathbf{z}^P_{k,m},\widetilde{\mathbf{z}}^P_{k',m'}\right)} \quad (4)$$

where $|\mathcal{B}_C|$ and $|\mathcal{B}_A|$ denote the numbers of clear and boundary-ambiguous instances, and (k', m') denotes another entity-modality instance. $s_E(\mathbf{x},\mathbf{y}) = e^{-D_E(\mathbf{x},\mathbf{y})/\tau_E}$ and $s_P(\mathbf{x},\mathbf{y}) = e^{-D_P(\mathbf{x},\mathbf{y})/\tau_P}$ convert the Euclidean and Poincaré distances D_E and D_P into similarities, with temperatures $\tau_E, \tau_P > 0$. $\mathcal{P}^{(E,P),(+,-)}_{k,m}$ denotes the positive and negative sample sets in the corresponding space. More detailed explanations and derivations are provided in Appendix D. $\mathbf{z}^P_{k,m}$ captures the relative geometric relation of each boundary-ambiguous instance to the clear low- and high-frequency references. We map it back to the tangent space $\bar{\mathbf{z}}^P_{k,m} = \log_0^c\left(\mathbf{z}^P_{k,m}\right)$ and use it as ambiguity-aware context to refine the original wavelet components. Specifically, the low-frequency component is reconstructed by jointly encoding its original wavelet representation and the learned ambiguity-aware context, while each high-frequency component is reconstructed independently to preserve its original wavelet scale.

$$\hat{\mathbf{c}}^L_{k,m} = \begin{cases}\mathbf{c}^L_{k,m}, \\ \mathbf{W}^L_D\left[\mathbf{c}^L_{k,m}\|\bar{\mathbf{z}}^P_{k,m}\right],\end{cases} \qquad \hat{\mathbf{c}}^{H,j}_{k,m} = \begin{cases}\mathbf{c}^{H,j}_{k,m}, & (k,m)\in\mathcal{B}_C \\ \mathbf{W}^{H,j}_D\left[\mathbf{c}^{H,j}_{k,m}\|\bar{\mathbf{z}}^P_{k,m}\right], & (k,m)\in\mathcal{B}_A\end{cases} \quad (5)$$

where $\mathbf{W}^L_D$ and $\mathbf{W}^{H,j}_D$ are learnable reconstruction matrices for the low-frequency branch and the j-th high-frequency scale, respectively. The original wavelet components preserve the underlying frequency-scale structure, while the Poincaré representation provides ambiguity-aware information for reconstructing boundary-ambiguous instances. For clear instances, the wavelet representations are retained, with their underlying features optimized by the Euclidean contrastive objective through end-to-end learning. For ambiguous instances, the corresponding low-frequency and multi-scale high-frequency representations are reconstructed under the learned Poincaré geometry. Consequently, $\widehat{\mathbf{c}}^L_{k,m}$ and $\{\widehat{\mathbf{c}}^{H,j}_{k,m}\}_{j=1}^J$ serve as the disambiguated low-frequency and multi-scale high-frequency representations for subsequent differentiated processing.

3.4 High-Frequency Relation-Aware Gating Strategy

HRGS applies relation-aware gating to the high-frequency representations disambiguated by DSFD, first across scales within each modality and then across structural, visual, and textual modalities, emphasizing relation-relevant cues. Meanwhile, the selected high-frequency representations from each modality provide discriminative guidance for low-frequency shared-memory interaction in LSIS. We project the j-th disambiguated high-frequency component into a unified space by $\mathbf{z}^{H,j}_{k,m} = \mathbf{W}^m_{H,j}\widehat{\mathbf{c}}^{H,j}_{k,m}$, where $\mathbf{W}^m_{H,j}$ is the learnable projection matrix for modality m and scale j. Subsequently, the current relation semantics are leveraged to compute the contribution of each high-frequency scale:

$$a^r_{k,m,j} = \mathbf{w}_\rho^\top \tanh\left(\mathbf{W}_\rho\left[\mathbf{z}^{H,j}_{k,m}\|\mathbf{e}_r\right]\right), \rho^r_{k,m,j} = \mathrm{softmax}_j\left(a^r_{k,m,j}\right), \widetilde{\mathbf{z}}^H_{k,m} = \textstyle\sum_{j=1}^J \rho^r_{k,m,j}\mathbf{z}^{H,j}_{k,m}, \quad (6)$$

where $a^r_{k,m,j}$ denotes the unnormalized relevance score of the j-th high-frequency scale in modality m of entity e_k under relation r, $\mathbf{W}_\rho$ and $\mathbf{w}_\rho$ are a learnable projection matrix and vector, respectively, and $\rho^r_{k,m,j}$ is the soft gating weight of scale j. The disambiguated multi-scale high-frequency representations are adaptively selected according to the current relation to obtain $\widetilde{\mathbf{z}}^H_{k,m}$. Using the relation-aware high-frequency representations from the three modalities, we construct the high-frequency guidance signal $\mathbf{G}^H_k = \mathbf{W}_G[\widetilde{\mathbf{z}}^H_{k,s}\|\widetilde{\mathbf{z}}^H_{k,v}\|\widetilde{\mathbf{z}}^H_{k,t}\|\mathbf{e}_r]$, where $\mathbf{W}_G$ is a learnable projection matrix. $\mathbf{G}^H_k$ retains the selected high-frequency information from the three modalities and incorporates relation semantics, serving specifically to control memory communication in LSIS, thereby avoiding modality bias during low-frequency shared communication.

Following intra-modality scale selection, the three modalities still contribute differently across relations, necessitating cross-modality selection:

$$b^H_{k,m} = (\mathbf{w}^H_\beta)^\top \tanh\big(\mathbf{W}^H_\beta[\tilde{\mathbf{z}}^H_{k,m} \| \mathbf{e}_r]\big),\ \beta^H_{k,m} = \mathrm{softmax}_m(b^H_{k,m}),\ \mathbf{F}^H_k = \sum \beta^H_{k,m}\tilde{\mathbf{z}}^H_{k,m}. \quad (7)$$

Here, $b^H_{k,m}$ denotes the unnormalized contribution score of modality m of entity e_k to the final high-frequency representation under relation r, $\mathbf{W}^H_\beta$ and $\mathbf{w}^H_\beta$ are a learnable matrix and vector, respectively. $\beta^H_{k,m}$ denotes the contribution of modality m under relation r. $\mathbf{F}^H_k$ is the final high-frequency representation after gated fusion of the three modalities.

3.5 Low-Frequency Shared-Memory Interaction Strategy

To better facilitate shared feature exchange and complementary semantic accumulation across different modalities, LSIS uses the disambiguated low-frequency representations from DSFD for cross-modal communication, where modality-specific communicators jointly access a hierarchical shared memory bank. LSIS enables cross-modal communication through a shared memory $\mathcal{M}^L = \{\mathcal{M}^g, \mathcal{M}^r, \mathcal{M}^{tr}\}$, corresponding to Global low-frequency Prototype Memory, Relation-conditioned low-frequency Memory, and Cross-modal Communication Trace Memory, which store stable shared semantics, relation-relevant semantics, and cross-modal interaction traces, respectively, with N key-value slots $(\mathbf{k}^\ell_n, \mathbf{v}^\ell_n)$ in each bank $\ell \in \{g, r, tr\}$. For entity e_k and modality m, the communicator receives $\mathbf{z}^L_{k,m} = \mathbf{W}^m_L \hat{\mathbf{c}}^L_{k,m}$, obtained using the modality-specific projection matrix $\mathbf{W}^m_L$. At training step t, communicators read from $\mathcal{M}^L_{t-1}$ and then write the current mini-batch contents to obtain $\mathcal{M}^L_t$. Bank design details and related experiments are provided in Appendix E.

For memory writing, the write gate of each modality-specific communicator controls how much current low-frequency information can be shared:

$$\mathbf{g}^w_{k,m} = \sigma\left(\mathbf{W}_w[\mathbf{z}^L_{k,m} \| \mathbf{G}^H_k \| \mathbf{e}_r]\right), \qquad \mathbf{w}^L_{k,m} = \mathbf{g}^w_{k,m} \odot \mathbf{W}^m_{mem}\mathbf{z}^L_{k,m}. \quad (8)$$

Here, $\mathbf{g}^w_{k,m}$ is the write gate, and $\mathbf{W}_w$ is learnable. The current low-frequency representation is then projected by $\mathbf{W}^m_{mem}$ and filtered by $\mathbf{g}^w_{k,m}$ to obtain the shareable content $\mathbf{w}^L_{k,m}$. The communicator further estimates its hierarchical preference over the three memory levels as $\xi^\ell_{k,m} = \mathrm{softmax}_\ell\left((\mathbf{w}^\ell_\xi)^\top[\mathbf{w}^L_{k,m} \| \mathbf{e}_r \| \mathbf{G}^H_k]\right), \ell \in \{g, r, tr\}$, where $\mathbf{w}^\ell_\xi$ is the learnable routing parameter and $\xi^\ell_{k,m}$ represents the hierarchical preference of communicator m over memory level ℓ, with $\sum_{\ell\in\{g,r,tr\}} \xi^\ell_{k,m} = 1$. The same hierarchical preference is used to route the shareable content during memory writing and to weight the corresponding level-specific retrieval during memory reading. Within each memory level, the routed content is further assigned to individual memory slots according to key-based addressing, $\eta^\ell_{k,m,n} = \mathrm{softmax}_n\left(\frac{(\mathbf{w}^L_{k,m})^\top \mathbf{k}^\ell_n}{\sqrt{d}}\right), n = 1, \ldots, N$, where $\eta^\ell_{k,m,n}$ denotes the write weight of the current shareable content to the n-th slot at memory level ℓ. These write weights determine how the shareable low-frequency content will be assigned to the hierarchical memory slots, while the actual memory values are updated after memory retrieval.

For memory reading, each communicator constructs a context-aware query that reflects its current representation and retrieval demand: $\mathbf{Q}_{k,m} = \mathbf{W}_Q[\mathbf{z}^L_{k,m} \| \mathbf{G}^H_k \| \mathbf{e}_r]$. Here, $\mathbf{Q}_{k,m}$ denotes the query of modality m, and $\mathbf{W}_Q$ is a learnable projection matrix. For each memory level ℓ, the query is matched against its keys, where $\mathbf{Q}^\top_{k,m}\mathbf{k}^\ell_n$ measures the relevance between the current query and the n-th memory slot. The corresponding attention weight is computed as:

$$\pi^\ell_{k,m,n} = \mathrm{softmax}_n\left(\frac{\mathbf{Q}^\top_{k,m}\mathbf{k}^\ell_n}{\sqrt{d}}\right), \qquad \mathbf{R}^\ell_{k,m} = \textstyle\sum_{n=1}^N \pi^\ell_{k,m,n}\mathbf{v}^\ell_{n,t-1}, \qquad \ell \in \{g, r, tr\}, \quad (9)$$

where d is the memory representation dimension, $\pi^\ell_{k,m,n}$ denotes the attention weight assigned to the n-th slot at level ℓ, and $\sum_{n=1}^N \pi^\ell_{k,m,n} = 1$. The retrieved low-frequency information $\mathbf{R}^\ell_{k,m}$ is obtained by weighted aggregation of the corresponding values $\mathbf{v}^\ell_{n,t-1}$ according to $\pi^\ell_{k,m,n}$. The level-specific retrievals are further aggregated according to the same hierarchical memory preference, $\mathbf{R}_{k,m} = \sum_{\ell\in\{g,r,tr\}} \xi^\ell_{k,m}\mathbf{R}^\ell_{k,m}$. Here, $\xi^\ell_{k,m}$ determines the contribution of memory level ℓ to the final retrieval, while $\pi^\ell_{k,m,n}$ determines the relevance of individual slots within that level.

For memory updating, the retrieved information is incorporated into the current modality through an update gate:

$$\mathbf{g}^u_{k,m} = \sigma\left(\mathbf{W}_u[\mathbf{z}^L_{k,m} \| \mathbf{R}_{k,m} \| \mathbf{G}^H_k \| \mathbf{e}_r]\right), \qquad \mathbf{u}^L_{k,m} = \mathbf{z}^L_{k,m} + \mathbf{g}^u_{k,m} \odot \mathbf{R}_{k,m}. \tag{10}$$

Here, $\mathbf{W}_u$ is learnable and $\mathbf{g}^u_{k,m}$ controls how much retrieved information is incorporated. $\mathbf{u}^L_{k,m}$ is the low-frequency representation after shared-memory updating. Then, the shareable contents of the current mini-batch are written back to update the persistent memory values,

$$\mathbf{v}^\ell_{n,t} = (1-\mu)\mathbf{v}^\ell_{n,t-1} + \mu \frac{\sum_{(k,m)\in\mathcal{I}_t} \xi^\ell_{k,m}\eta^\ell_{k,m,n}\mathbf{w}^L_{k,m}}{\sum_{(k,m)\in\mathcal{I}_t} \xi^\ell_{k,m}\eta^\ell_{k,m,n} + \epsilon}, \qquad \ell \in \{g, r, tr\}, \tag{11}$$

where $\mathcal{I}_t$ denotes the entity-modality instances in the current mini-batch, μ is the memory update rate, and ϵ is a small constant for numerical stability. The previous memory state $\mathbf{v}^\ell_{n,t-1}$ preserves accumulated shared semantics, while the newly routed low-frequency contents progressively update the memory for subsequent mini-batches. The memory keys remain stable for addressing during this process. After shared-memory interaction, the updated low-frequency representations are fused using relation-conditioned modality weights: $\mathbf{F}^L_k = \sum_{m\in\{s,v,t\}} \beta^L_{k,m}\mathbf{u}^L_{k,m}$. Here, $\beta^L_{k,m}$ is the normalized weight of modality m, computed using the same relation-conditioned scoring form as Eq. equation 7, with independent low-frequency parameters. The full formulation is provided in Appendix E.3. The final shared low-frequency representation $\mathbf{F}^L_k$ is obtained by weighting $\mathbf{u}^L_{k,m}$ with $\beta^L_{k,m}$. Finally, the low- and high-frequency representations are adaptively fused as $\mathbf{H}^{out}_k = \gamma_k\mathbf{F}^L_k + (1-\gamma_k)\mathbf{F}^H_k$, where $\gamma_k = \sigma\big(\mathbf{w}^\top_\gamma \left[\mathbf{F}^L_k \| \mathbf{F}^H_k \| \mathbf{e}_r\right]\big)$, controls their contributions. The fused entity representation $\mathbf{H}^{out}_k$ is then directly used for triple scoring as $S(h,r,t) = \phi(\mathbf{H}^{out}_h, \mathbf{e}_r, \mathbf{H}^{out}_t)$. The overall model is jointly optimized by the MMKGC prediction objective and the auxiliary dual-space frequency disambiguation objective: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{KGC}} + \lambda_{\text{DSFD}}\mathcal{L}_{\text{DSFD}}$, where $\mathcal{L}_{\text{KGC}}$ denotes the link prediction loss and λ_{DSFD} controls the contribution of the DSFD objective.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct our experiments on three public MMKG benchmarks: MKG-W (Xu et al., 2022), MKG-Y (Xu et al., 2022), and DB15K (Liu et al., 2019). Detailed statistics of the datasets are shown in Appendix F.

Baseline Methods. To make comprehensive comparisons, we choose representative recent KGC methods as baselines. The first category is MMKGC methods, including IMF (Li et al., 2023), VISTA (Lee et al., 2023), AdaMF-MAT (Zhang et al., 2024b), NativE (Zhang et al., 2024a), MoMoK (Zhang et al., 2025), MCKGC (Gao et al., 2025), DHNS (Niu & Zhang, 2025), and M-Hyper (Liu et al., 2026b). The second category is LLM-based MMKGC methods, including SNAG (Chen et al., 2025), HERGC (Xiao & Zhang, 2026), and HFR-MKGC (Wang et al., 2026). The third category is frequency-domain MMKGC methods, including WFF (Xu et al., 2025) and MLFormer (Wang et al., 2025a).

Implementation Details. DSF^2 is optimized with Adam on an NVIDIA RTX 4090 GPU, with a batch size of 1000, a learning rate of 10^{-3}, dropout of 0.3, and 3 graph layers. We use $K = 20$ for modality-specific graphs, $J = 3$ wavelet scales, 24 memory slots per LSIS level, and boundary threshold $\delta = 0.20$. More implementation details are provided in Appendix F.2.

4.2 Main Results

The main MMKGC results are reported in Table 1. DSF^2 achieves the best MRR, Hits@1, and Hits@3 on all datasets, while also obtaining the best Hits@10 on DB15K. It improves the second-best MRR by 0.83, 0.49, and 1.50 points and the MoMoK backbone by 11.90%, 8.06%, and 18.17%, respectively. These results indicate that DSF^2 is particularly effective at ranking correct entities among the top candidates, benefiting from the differentiated modeling of low-frequency shared semantics and high-frequency relation-discriminative information. Stability and statistical significance over 10 independent runs are reported in Appendix F.3.

Table 1: Main results. Best results are shown in **bold**, and second-best results are underlined. $*$ denotes results reported in the original papers, and † denotes results reproduced using public code. MoMoK (2025) serves as our backbone.

	MKG-W				MKG-Y				DB15K			
Models	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
Multimodal KGC												
IMF(23')*	34.50	28.77	36.62	45.44	35.79	32.95	37.14	40.63	32.25	24.20	36.00	48.19
VISTA(23')*	32.91	26.12	35.38	45.61	30.45	24.87	32.39	41.53	30.42	22.49	33.56	45.94
AdaMF-MAT(24')*	35.85	29.04	39.01	48.42	38.57	34.34	40.59	45.76	35.14	25.30	41.11	52.92
NativE(24')†	37.20	30.40	40.47	50.28	39.00	34.87	40.47	46.24	36.85	27.81	41.31	53.93
MoMoK(25')†	35.70	30.47	37.68	46.14	37.71	34.81	39.32	42.98	39.52	32.00	43.49	54.51
MCKGC(25')*	36.88	31.32	38.92	47.43	38.92	35.49	40.57	45.21	39.79	31.92	43.80	54.66
DHNS(25')*	35.68	29.98	38.68	48.12	39.11	34.70	41.23	46.66	34.36	23.34	41.56	53.72
M-Hyper(26')†	37.12	31.16	38.96	48.67	39.48	36.00	40.62	45.25	41.35	33.54	44.61	55.79
LLM-based MMKGC												
SNAG(25')*	37.30	30.20	40.50	50.30	39.50	35.40	41.10	47.10	36.30	27.40	41.10	53.00
HFR-MKGC(26')*	38.62	31.37	42.12	**51.66**	39.00	34.41	42.12	**47.31**	35.20	23.04	43.25	56.41
HERGC(26')*	39.12	33.65	41.80	48.54	40.26	36.73	42.22	45.22	40.95	33.47	45.66	55.38
Frequency-domain MMKGC												
WFF(25')†	37.07	29.06	38.17	46.16	39.56	35.03	40.50	43.89	45.20	38.65	48.58	58.09
MLFormer(25')†	34.79	29.63	36.36	44.71	38.03	34.90	39.47	43.52	36.57	29.26	39.96	50.71
DSF^2 (Ours)	**39.95**	**34.45**	**43.65**	51.20	**40.75**	**37.05**	**43.55**	46.05	**46.70**	**40.00**	**50.20**	**59.50**

Frequency-domain methods show a clear advantage on DB15K, whose richer relation types and denser graph connectivity provide informative structural patterns for modeling smooth shared semantics and relation-specific local variations, which semantic enhancement relying mainly on large pretrained models struggles to fully capture. LLM-based methods achieve several best or second-best results on MKG-W and MKG-Y, but their performance drops on DB15K, with some metrics below those of conventional multimodal KGC methods. DSF^2 outperforms most LLM-based results on MKG-W and MKG-Y and improves over the frequency-domain baseline on DB15K, demonstrating its ability to model both shared smooth semantics and local discriminative variations.

4.3 Ablation Study

To evaluate the contribution of each component, we conduct ablation studies on three datasets. The experimental results are presented in Table 2. Removing any component consistently degrades performance. Removing LF-Inter and HF-Selection leads to clear performance degradation, demonstrating the importance of cross-modal low-frequency semantic accumulation and relation-discriminative high-frequency selection. Removing DSFD decreases the average MRR by 2.08 points, validating dual-space modeling for clear and ambiguous entity-modality instances. The performance drops caused by removing high-frequency guidance further confirm the necessity of high-frequency-guided low-frequency interaction.

Table 2: Ablation results on the three datasets. The best results are shown in **bold**.

	MKG-W		MKG-Y		DB15K	
Variant	MRR	H@1	MRR	H@1	MRR	H@1
DSF^2 (Ours)	**39.95**	**34.45**	**40.75**	**37.05**	**46.70**	**40.00**
w/o Multi-Graph	38.21	32.78	39.06	35.34	44.91	37.12
w/o DSFD	37.86	32.31	38.74	34.96	44.55	36.71
w/o LF-Inter	37.42	31.88	38.31	34.57	43.96	36.12
w/o HF-Selection	37.68	32.05	38.52	34.72	44.18	36.39
w/o HF-Guidance	38.05	32.64	38.93	35.20	44.72	36.88

4.4 Differentiated Frequency Processing

To further examine whether low and high-frequency representations require different processing strategies by exchanging shared-memory communication and relation-aware selection between the two branches. As shown in Figure 3. Our differentiated strategy achieves the best overall performance, while exchanging or unifying the processing strategies causes clear degradation. This validates differentiated frequency processing, with stable low frequency semantics favoring cross-modal memory communication and high-frequency variations benefiting from relation-aware selection.

4.5 Effect of Space Mapping Strategies in DSF^2

We further investigate the effect of different dual-space mapping strategies. As shown in Table 3, *Euc/Poi* achieves the best performance on all three datasets. Compared with *Euc/Euc*, it improves MRR by 1.53, 1.57, and 1.58 points on three datasets. In contrast, reversing the mapping leads to the largest degradation, demonstrating that Euclidean space is more suitable for modeling clear-frequency representations, whereas Poincaré geometry provides a flexible non-Euclidean metric structure for modeling representations of entity–modality instances with ambiguous frequency dominance relative to clear-frequency references.

Table 3: Performance comparison of different frequency-space mapping strategies. The best results are shown in **bold**.

Clear/ Ambiguous	MKG-W		MKG-Y		DB15K	
	MRR	H@1	MRR	H@1	MRR	H@1
Euc / Euc	38.42	33.01	39.18	35.92	45.12	37.42
Poi / Poi	37.86	32.54	38.61	35.31	44.63	36.81
Poi / Euc	37.21	31.82	37.94	34.66	43.82	36.02
Euc / Poi	**39.95**	**34.45**	**40.75**	**37.05**	**46.70**	**40.00**

4.6 Effect of Shared Memory

To explore whether memory sharing is essential for low frequency cross modal interaction, we compare three memory configurations with other components fixed, *No Memory* directly fuses low-frequency representations, *Private Memory* maintains independent memories $\mathbf{M}_s^L$, $\mathbf{M}_v^L$, and $\mathbf{M}_t^L$ for the three modalities, and *Shared Memory* lets all communicators access $\mathbf{M}^L$. As depicted in Figure 4, confirming the benefits of memory-based semantic accumulation and the additional gains from explicit cross-modal low-frequency sharing.

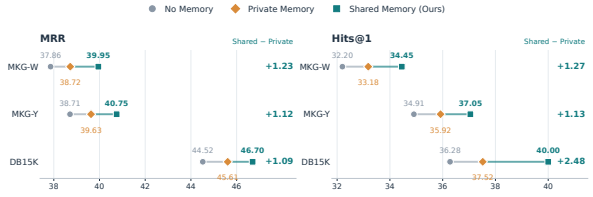


Figure 4: Comparison of different memory configurations.

4.7 Hyperparameter Analysis

Across all three datasets, performance peaks at boundary threshold $\delta = 0.20$ (Figure 5(a)), smaller values may classify ambiguous frequencies as clear, whereas larger values may weaken frequency discrimination. Increasing the number of high-frequency scales improves performance up to $J = 3$, beyond which redundant or noisy high-frequency information causes a slight decline (Figure 5(b)).

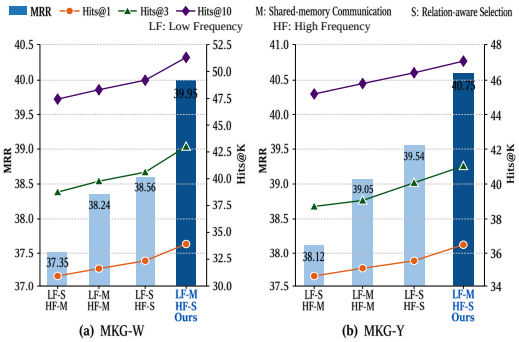


Figure 3: Low- and high-frequency strategy comparison. Labels: *(frequency component)-(strategy)*.

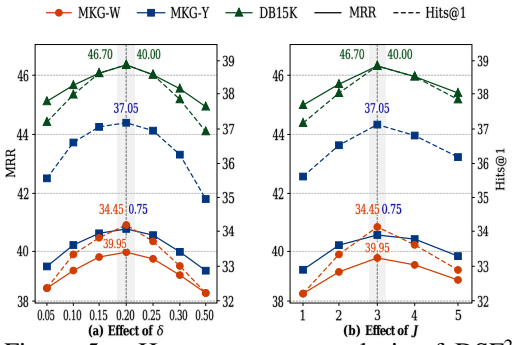


Figure 5: Hyperparameter analysis of DSF^2: ambiguous-band threshold δ and high-frequency scale count J.

5 Conclusion

In DSF^2, dual-space disambiguation is integrated with differentiated frequency processing for MMKGC. Ambiguous Poincaré representations are refined against clear Euclidean references in DSFD, while stable low-frequency semantics are accumulated across modalities through shared memory in LSIS, guided by relation-discriminative high-frequency cues selected by HRGS. Improvements on three benchmarks demonstrate its effectiveness.

AI USE STATEMENT

AI tools were utilized during the preparation of this paper. Specifically, AI tools were used for grammar correction and wording polishing of manuscript text, as well as for information retrieval, identification of relevant studies, and literature searches.

The authors take full responsibility for all technical insights, model designs, experimental analyses, and the final manuscript. During manuscript preparation, AI assistance was used to refine grammar, wording, terminology, clarity, and organization, as well as to assist with translation and related paper search. The authors take full responsibility for the originality, accuracy, and integrity of the final submission, including its text, references, methods, code, analyses, and reported results.

REFERENCES

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. Knowledge graphs meet multi-modal learning: A comprehensive survey. *CoRR*, abs/2402.05391, 2024. doi: 10.48550/arXiv.2402.05391.

Zhuo Chen, Yin Fang, Yichi Zhang, Lingbing Guo, Jiaoyan Chen, Jeff Z. Pan, Huajun Chen, and Wen Zhang. Noise-powered multi-modal knowledge graph representation framework. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 141–155, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.11/.

Jinman Cui, Dongmei Li, Xinyang Wang, Lei Yao, Yakun Li, Xiaolong Qu, and Wenbin Zhang. Modality-aware instruction tuning of large language models for multimodal knowledge graph completion. *Information Fusion*, 134:104387, 2026. doi: 10.1016/j.inffus.2026.104387.

Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. VL-KGE: Vision–language models meet knowledge graph embeddings. In *Proceedings of the ACM Web Conference 2026*, pp. 7552–7563. Association for Computing Machinery, 2026. doi: 10.1145/3774904.3792677.

Yuxiao Gao, Fuwei Zhang, Zhao Zhang, Xiaoshuang Min, and Fuzhen Zhuang. Mixed-curvature multi-modal knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11699–11707, 2025. doi: 10.1609/aaai.v39i11.33273.

David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011. doi: 10.1016/j.acha.2010.04.005.

Yue Jian, Xiangyu Luo, Zhifei Li, Miao Zhang, Yan Zhang, Kui Xiao, and Xiaoju Hou. APKGC: Noise-enhanced multi-modal knowledge graph completion with attention penalty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15005–15013, 2025. doi: 10.1609/aaai.v39i14.33645.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. VISTA: Visual-textual knowledge graph representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7314–7328. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.488.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6 (2):167–195, 2015. doi: 10.3233/SW-140134.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. IMF: Interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, pp. 2572–2580. Association for Computing Machinery, 2023. doi: 10.1145/3543507.3583554.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (12):9456–9478, 2024. doi: 10.1109/TPAMI.2024.3417451.

Haiyan Liu, Xin Song, Ye Wang, Kai Chen, Yuying Liu, Bin Zhou, Hongkui Tu, Liqun Gao, and Yue Qian. Robust multi-modal knowledge graph completion via modality-specific experts. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 728–737. Association for Computing Machinery, 2026a. doi: 10.1145/3805622.3810765.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. MMKG: Multi-modal knowledge graphs. In *The Semantic Web – 16th International Conference, ESWC 2019*, pp. 459–474. Springer, 2019. doi: 10.1007/978-3-030-21348-0_30.

Zhiqiang Liu, Yichi Zhang, Mengshu Sun, Lei Liang, and Wen Zhang. Collaboration of fusion and independence: Hypercomplex-driven robust multi-modal knowledge graph completion. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27968–27983. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.1289.

Hung Nguyen, Ban Tran, and Thanh Le. Advancing multimodal knowledge graphs with KAN fusion and quaternion Fourier transform filtering. In *Advances in Computational Collective Intelligence*, volume 2747 of *Communications in Computer and Information Science*, pp. 324–338. Springer, 2026. doi: 10.1007/978-3-032-10202-7_22.

Guanglin Niu and Xiaowei Zhang. Diffusion-based hierarchical negative sampling for multimodal knowledge graph completion. In *Database Systems for Advanced Applications*, pp. 479–495. Springer, 2025. doi: 10.1007/978-981-95-3906-2_34.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pp. 697–706. Association for Computing Machinery, 2007. doi: 10.1145/1242572.1242667.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489.

Thanh Vu, Huy Ngo, Bac Le, and Thanh Le. FouriER: Link prediction by mixing tokens with fourier-enhanced metaformer. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2023. doi: 10.14428/esann/2023.ES2023-73.

Di Wang, Junping Du, Zhe Xue, Meiyu Liang, Guanhua Ye, Yingxia Shao, and Haisheng Li. HFR-MKGC: Hierarchical fusion reasoning with MLLMs for multi-modal knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15815–15823, 2026. doi: 10.1609/aaai.v40i18.38613.

Meng Wang, Changyu Li, Feiyu Chen, Jie Shao, Ke Qin, and Shuang Liang. MLFormer: Unleashing efficiency without attention for multimodal knowledge graph embedding. *IEEE Transactions on Computational Social Systems*, 12(6):5536–5549, 2025a. doi: 10.1109/TCSS.2025.3620089.

Zheng Wang, Taiyu Li, and Zengzhao Chen. RASD: Relation aware spectral decoupling attention network for knowledge graph reasoning. *Applied Sciences*, 15(16):9049, 2025b. doi: 10.3390/app15169049.

Feifei Wei and Kuizhi Mei. Frequency inception based graph neural network for relation prediction in knowledge graphs. *Knowledge-Based Systems*, 278:110908, 2023. doi: 10.1016/j.knosys.2023.110908.

Yongkang Xiao and Rui Zhang. HERGC: Heterogeneous experts representation and generative completion for multimodal knowledge graphs. In *Advances in Knowledge Discovery and Data Mining*, pp. 327–339. Springer, 2026. doi: 10.1007/978-981-92-1300-9_26.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3857–3866. Association for Computing Machinery, 2022. doi: 10.1145/3503161.3548388.

Xiaodi Xu, Lijie Li, Ye Wang, Tao Ren, and Tian Qiao. WFF: Wavelet-based information fusion for multimodal knowledge graph link prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2084–2093. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3755661.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. NativE: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 91–101. Association for Computing Machinery, 2024a. doi: 10.1145/3626772.3657800.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17120–17130, Torino, Italia, 2024b. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1487/.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. In *International Conference on Learning Representations*, 2025.

A Related Work

A.1 MMKGC Studies

Multimodal Knowledge Graph Completion (MMKGC) aims to infer missing facts in a knowledge graph by jointly exploiting structural information and the multimodal content associated with entities. Existing MMKGC methods mainly operate in the representation space and can be grouped according to their prediction mechanisms.

Discriminative completion. IMF (Li et al., 2023) introduces a two-stage fusion framework that combines bilinear interaction, contrastive constraints, and decision-level fusion to exploit cross-modal complementarity. APKGC (Jian et al., 2025) improves robustness through attention regularization and adaptive noise augmentation, whereas MEIF (Liu et al., 2026a) uses modality-specific experts and adaptive routing to handle incomplete modalities. NativE (Zhang et al., 2024a) employs relation-guided adaptive fusion to address modality diversity and missing modalities. MoMoK (Zhang et al., 2025) introduces relation-guided modality experts for complementary knowledge modeling. M-Hyper (Liu et al., 2026b) integrates independent and fused modality representations through hypercomplex modeling.

Generative and hybrid completion. SNAG (Chen et al., 2025) improves multimodal representation robustness through modality-level noise masking. HFR-MKGC (Wang et al., 2026) combines hierarchical multimodal fusion with MLLM-based reasoning. HERGC (Xiao & Zhang, 2026) leverages heterogeneous experts and LLM-based generative prediction for MMKGC.

A.2 Frequency-Domain Studies

Recent frequency-domain studies can be divided into two categories.

Frequency-domain methods for MMKGC. WFF (Xu et al., 2025) applies discrete wavelet transforms to frequency decomposition and cross-modal fusion. MLFormer (Wang et al., 2025a) combines filter gating with Fourier transforms to improve multimodal representation learning. MMQuat (Nguyen et al., 2026) integrates KAN fusion with quaternion Fourier filtering to suppress multimodal noise.

Frequency-domain methods for KGC. FiGNN (Wei & Mei, 2023) performs frequency-aware graph modeling for relation prediction. FouriER (Vu et al., 2023) adopts Fourier-based token mixing to model entity–relation interactions for link prediction. RASD (Wang et al., 2025b) combines spectral decoupling attention with relation-aware filtering for knowledge graph reasoning.

B Multimodal Feature Encoding and Graph Wavelet Decomposition

Modality-specific feature encoding. For the structural modality, we maintain a trainable entity embedding matrix. Its rows are randomly initialized and optimized jointly with the model using only the observed training triples. For the visual and textual modalities, pretrained VGG and BERT encoders are used to extract features from entity-associated images and textual descriptions, respectively. The features of all entities are stacked in a consistent entity order:

$$\mathbf{X}_m = \begin{bmatrix} \mathbf{x}_{1,m}^\top \\ \vdots \\ \mathbf{x}_{|\mathcal{E}|,m}^\top \end{bmatrix} \in \mathbb{R}^{|\mathcal{E}| \times d_m}. \tag{12}$$

The same row index identifies the same entity across all three modalities, therefore, the feature matrices and modality-specific adjacency matrices use a consistent node ordering. Here, $\mathbf{X}_s$ contains trainable structural embeddings, whereas $\mathbf{X}_v$ and $\mathbf{X}_t$ contain visual and textual encoder features. Because the input dimensions can differ across modalities, we apply a modality-specific projection

followed by layer normalization:

$$\mathbf{H}_m = \mathrm{LN}(\mathbf{X}_m \mathbf{W}_m^{\mathrm{enc}}) \in \mathbb{R}^{|\mathcal{E}| \times d}. \tag{13}$$

Here, $\mathbf{W}_m^{\mathrm{enc}} \in \mathbb{R}^{d_m \times d}$ is a learnable projection, d is the shared output dimension, and LN denotes layer normalization across the feature channels of each entity. The resulting $\mathbf{H}_m$ is treated as a graph signal on $\mathcal{G}_m$: its rows correspond to entities and its columns correspond to feature channels.

Graph Laplacian and spectral graph wavelet decomposition. For modality $m \in \{s, v, t\}$, the graph $\mathcal{G}_m = (\mathcal{E}, \mathbf{A}_m)$ defines the neighborhood structure over the shared entity set. Let $n = |\mathcal{E}|$. The adjacency matrix supplied to the spectral operator is symmetric and nonnegative, i.e., $\mathbf{A}_m \in \mathbb{R}_{\geq 0}^{n \times n}$. Its diagonal degree matrix $\mathbf{D}_m$ is

$$(\mathbf{D}_m)_{ii} = \sum_{j=1}^{n} (\mathbf{A}_m)_{ij}. \tag{14}$$

The symmetric normalized graph Laplacian is

$$\mathbf{L}_m = \mathbf{I} - \mathbf{D}_m^{-1/2} \mathbf{A}_m \mathbf{D}_m^{-1/2}. \tag{15}$$

Since $\mathbf{L}_m$ is real and symmetric, it admits the orthonormal eigendecomposition

$$\mathbf{L}_m = \mathbf{U}_m \mathbf{\Lambda}_m \mathbf{U}_m^\top, \qquad \mathbf{U}_m^\top \mathbf{U}_m = \mathbf{I}. \tag{16}$$

Here, $\mathbf{U}_m = [\mathbf{u}_{m,1}, \ldots, \mathbf{u}_{m,n}]$ contains the eigenvectors and $\mathbf{\Lambda}_m = \mathrm{diag}(\lambda_{m,1}, \ldots, \lambda_{m,n})$ contains the eigenvalues. For a symmetric normalized Laplacian, $\lambda_{m,i} \in [0, 2]$. For a unit-norm eigenvector,

$$\mathbf{u}_{m,i}^\top \mathbf{L}_m \mathbf{u}_{m,i} = \lambda_{m,i}. \tag{17}$$

Thus, smaller eigenvalues correspond to smoother graph patterns, whereas larger eigenvalues correspond to greater graph variation. Each modality has its own frequency basis because its adjacency matrix defines a different neighborhood structure.

The encoded feature matrix $\mathbf{H}_m \in \mathbb{R}^{n \times d}$ contains d feature channels over the entities. Multiplication by $\mathbf{U}_m^\top$ expresses each channel in the graph Fourier basis. The spectral kernels are applied to the corresponding eigenvalues:

$$h(\mathbf{\Lambda}_m) = \mathrm{diag}(h(\lambda_{m,1}), \ldots, h(\lambda_{m,n}))\,, \tag{18}$$

$$g(s_j \mathbf{\Lambda}_m) = \mathrm{diag}(g(s_j \lambda_{m,1}), \ldots, g(s_j \lambda_{m,n}))\,. \tag{19}$$

Here, h is the low-frequency kernel and $g(s_j\,\cdot)$ is the wavelet kernel at scale j, where $j = 1, \ldots, J$. Multiplication by $\mathbf{U}_m$ maps the filtered signals back to entity coordinates and produces $\mathbf{C}_m^L$ and $\mathbf{C}_m^{H,j}$ as defined in Eq. equation 1. Both matrices have shape $n \times d$. Their k-th rows provide the low-frequency response $\mathbf{c}_{k,m}^L$ and the scale-specific high-frequency response $\mathbf{c}_{k,m}^{H,j}$ used by the subsequent modules.

Multi-scale wavelet responses. We use $J = 3$ wavelet scales, $(s_1, s_2, s_3) = (2, 1, 0.5)$. For the kernel specified in Appendix C, the response peak occurs at $\lambda_j^\star = 1/s_j$, giving peak locations 0.5, 1, and 2. Accordingly, $s_1 = 2$ emphasizes relatively coarse graph variation, $s_2 = 1$ captures intermediate variation, and $s_3 = 0.5$ emphasizes finer variation. Although these outputs are collectively called high-frequency representations, they are overlapping detail responses at different spectral scales rather than three disjoint bands. Retaining these responses separately allows HRGS to perform relation-conditioned scale selection.

C Dual-Space Frequency Disambiguation

C.1 Spectral Filter Responses and Transition-Band Analysis

To characterize the spectral overlap that motivates boundary disambiguation, we examine the responses of the low-pass and multi-scale wavelet filters. For $\lambda \in [0, 2]$, we use

$$h(\lambda) = e^{-4\lambda}, \qquad g(s_j \lambda) = s_j \lambda e^{1 - s_j \lambda}, \tag{20}$$

with $(s_1, s_2, s_3) = (2, 1, 0.5)$ and fixed energy-aggregation weights $\omega_j = 1/3$. The low-pass filter emphasizes smooth graph components. The wavelet filters suppress the zero-frequency component and attain their maximum responses at $\lambda = 0.5$, 1, and 2, respectively.

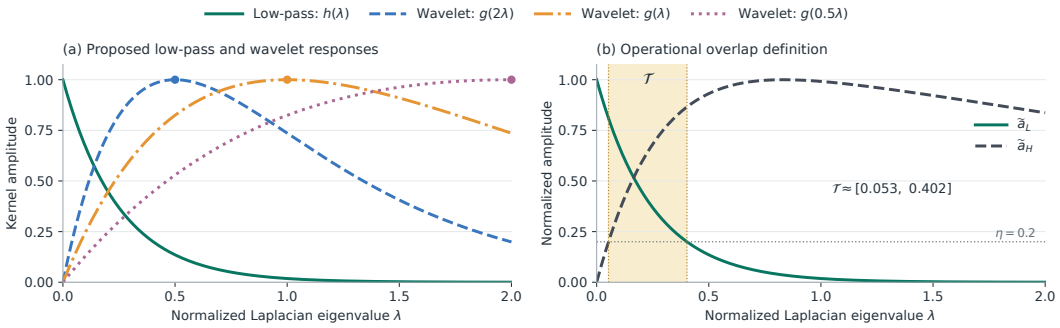


Figure 6: Spectral filter responses and the operational transition region. (a) The low-pass kernel and the three wavelet kernels with $s_j \in \{2, 1, 0.5\}$. (b) The normalized low-pass response and the aggregated wavelet response. The shaded interval is $\mathcal{T}_{0.2}$, where both normalized responses are at least 0.2.

We define the low- and high-frequency response amplitudes as

$$a^L(\lambda) = |h(\lambda)|, \qquad a^H(\lambda) = \left(\sum_{j=1}^{3} \omega_j |g(s_j \lambda)|^2 \right)^{1/2}. \tag{21}$$

After normalizing each response by its maximum on $[0, 2]$, denoted by $\widetilde{a}^L$ and $\widetilde{a}^H$, the transition region is

$$\mathcal{T}_\eta = \left\{ \lambda \in [0, 2] : \widetilde{a}^L(\lambda) \geq \eta, \widetilde{a}^H(\lambda) \geq \eta \right\}. \tag{22}$$

This normalization is used only to identify the transition region, it does not rescale the model branches or alter the energies used to compute $q_{k,m}$. With $\eta = 0.2$, numerical evaluation gives $\mathcal{T}_{0.2} \approx [0.053, 0.402]$. Figure 6 shows that both branches have appreciable responses in this interval, thereby providing an operational definition of the spectral overlap used in the following analyses.

C.2 Frequency Dominance and Transition-Band Responses

This analysis tests whether the energy-based criterion in Section 3.3 identifies instances with stronger transition-band responses. We fix the DSFD routing threshold at $\delta = 0.2$ and define

$$\begin{aligned} \mathcal{B}^L &= \{(k, m) : q_{k,m} < -0.2\}, \\ \mathcal{B}^A &= \{(k, m) : |q_{k,m}| \leq 0.2\}, \\ \mathcal{B}^H &= \{(k, m) : q_{k,m} > 0.2\}. \end{aligned} \tag{23}$$

The analysis uses the graph signals and wavelet responses immediately before DSFD. Model parameters and memory states are frozen, and each valid entity–modality pair is counted once. Structural, visual, and textual modalities are analyzed separately.

Transition-band response. The transition region is computed from the filters and effective aggregation weights used by the evaluated checkpoint. On a fixed dense grid $\mathcal{G} \subset [0, 2]$, let

$$a^L(\lambda) = |h(\lambda)|, \qquad a^H(\lambda) = \left(\sum_{j=1}^{J} \omega_j |g(s_j \lambda)|^2 \right)^{1/2}, \tag{24}$$

$$\widetilde{a}^r(\lambda) = \frac{a^r(\lambda)}{\max_{\xi \in \mathcal{G}} a^r(\xi)}, \qquad r \in \{L, H\}, \tag{25}$$

$$\mathcal{T}_\eta = \{\lambda : \widetilde{a}^L(\lambda) \geq \eta, \widetilde{a}^H(\lambda) \geq \eta\}. \tag{26}$$

We use $\eta = 0.2$ for the primary analysis and repeat the same analysis with $\eta \in \{0.1, 0.3\}$ as a sensitivity check. The threshold η identifies spectral overlap, whereas δ partitions instances according to frequency dominance. Neither the normalization nor η changes the model computation.

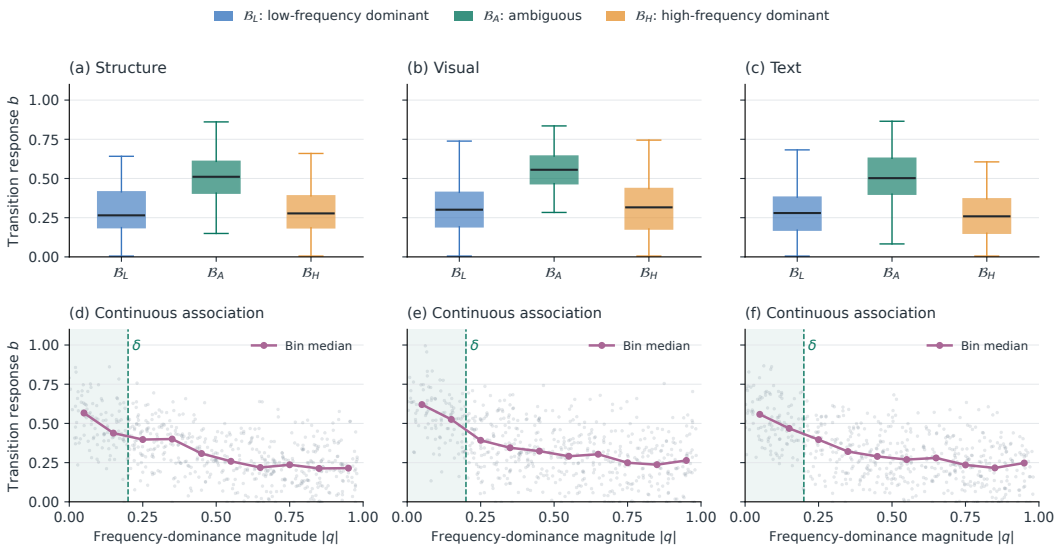


Figure 7: Relationship between frequency dominance and normalized transition response. (a)–(c) Distributions of b in $\mathcal{B}^L$, $\mathcal{B}^A$, and $\mathcal{B}^H$ for the structural, visual, and textual modalities. Boxes show interquartile ranges and center lines show medians. (d)–(f) Instance-level transition response against $|q|$, connected markers are medians in ten fixed-width bins. The dashed line marks the DSFD routing threshold $\delta = 0.2$.

For each modality, the transition-restricted responses are

$$\mathbf{C}^{L}_{m,\mathcal{T}} = \mathbf{U}_m \mathbf{1}_{\mathcal{T}_\eta}(\mathbf{\Lambda}_m) h(\mathbf{\Lambda}_m) \mathbf{U}_m^\top \mathbf{H}_m, \tag{27}$$

$$\mathbf{C}^{H,j}_{m,\mathcal{T}} = \mathbf{U}_m \mathbf{1}_{\mathcal{T}_\eta}(\mathbf{\Lambda}_m) g(s_j \mathbf{\Lambda}_m) \mathbf{U}_m^\top \mathbf{H}_m. \tag{28}$$

The complementary responses are $\overline{\mathbf{C}}^{L}_{m,\mathcal{T}} = \mathbf{C}^{L}_{m} - \mathbf{C}^{L}_{m,\mathcal{T}}$ and $\overline{\mathbf{C}}^{H,j}_{m,\mathcal{T}} = \mathbf{C}^{H,j}_{m} - \mathbf{C}^{H,j}_{m,\mathcal{T}}$. For each entity–modality pair, we compute

$$e^{\mathcal{T}}_{k,m} = \|\mathbf{c}^{L}_{k,m,\mathcal{T}}\|_2^2 + \sum_{j=1}^{J} \omega_j \|\mathbf{c}^{H,j}_{k,m,\mathcal{T}}\|_2^2, \tag{29}$$

$$b_{k,m} = \frac{e^{\mathcal{T}}_{k,m}}{e^{\mathcal{T}}_{k,m} + \bar{e}^{\mathcal{T}}_{k,m} + \epsilon_b}, \qquad \epsilon_b = 10^{-8}, \tag{30}$$

where $\bar{e}^{\mathcal{T}}_{k,m}$ is computed from the complementary responses. A larger $b_{k,m}$ indicates a stronger transition-band response relative to the complementary response. The definition of $b_{k,m}$ does not use $q_{k,m}$ or the resulting group assignment.

Results. Figure 7 shows that the ambiguous group $\mathcal{B}^A$ has a higher median normalized transition response than the two dominant groups for all three modalities. The binned medians generally decrease as $|q_{k,m}|$ increases, although local fluctuations and overlapping distributions remain. These results support the use of $q_{k,m}$ as a routing heuristic, but they do not imply that $q_{k,m}$ is a direct measurement of spectral coupling or that every ambiguous instance must have a large transition response.

C.3 DSFD Gains across Transition-Response Groups

This experiment measures whether the predictive benefit of DSFD varies with the strength of the transition-band response. For each dataset, we use the trained full DSF^2 model and its separately trained w/o-DSFD ablation. The two models are evaluated on identical filtered test queries with the same candidate sets and tie-handling policy. The checkpoint used to define the grouping is selected using validation performance only, model parameters, graph construction, filter kernels, and memory states are frozen during response extraction.

Using the filters and weights from this checkpoint, we compute

$$a^L(\lambda) = |h(\lambda)|, \qquad a^H(\lambda) = \left(\sum_{j=1}^{J} \omega_j |g(s_j \lambda)|^2 \right)^{1/2}, \tag{31}$$

$$\widetilde{a}^r(\lambda) = \frac{a^r(\lambda)}{\max_{\xi \in \mathcal{G}} a^r(\xi)}, \qquad r \in \{L, H\}, \tag{32}$$

$$\mathcal{T}_\eta = \{\lambda : \widetilde{a}^L(\lambda) \geq \eta, \widetilde{a}^H(\lambda) \geq \eta\}. \tag{33}$$

We use $\eta = 0.2$ for the primary analysis and repeat the evaluation with $\eta \in \{0.1, 0.3\}$ for sensitivity analysis. This analysis threshold is distinct from the DSFD routing threshold $\delta = 0.2$.

The transition-restricted responses are

$$\mathbf{C}^L_{m,\mathcal{T}} = \mathbf{U}_m \mathbf{1}_{\mathcal{T}_\eta}(\mathbf{\Lambda}_m) h(\mathbf{\Lambda}_m) \mathbf{U}_m^\top \mathbf{H}_m, \tag{34}$$

$$\mathbf{C}^{H,j}_{m,\mathcal{T}} = \mathbf{U}_m \mathbf{1}_{\mathcal{T}_\eta}(\mathbf{\Lambda}_m) g(s_j \mathbf{\Lambda}_m) \mathbf{U}_m^\top \mathbf{H}_m. \tag{35}$$

Complementary responses are obtained by subtracting the restricted responses from their corresponding original branches. For each entity–modality pair, we compute

$$\begin{aligned} e^{\mathcal{T}}_{k,m} &= \|\mathbf{c}^L_{k,m,\mathcal{T}}\|_2^2 + \sum_j \omega_j \|\mathbf{c}^{H,j}_{k,m,\mathcal{T}}\|_2^2, \\ b^{\mathrm{ref}}_{k,m} &= \frac{e^{\mathcal{T}}_{k,m}}{e^{\mathcal{T}}_{k,m} + \overline{e}^{\mathcal{T}}_{k,m} + 10^{-8}}, \\ b^{\mathrm{ref}}_k &= \frac{1}{|\mathcal{M}_k|} \sum_{m \in \mathcal{M}_k} b^{\mathrm{ref}}_{k,m}, \end{aligned} \tag{36}$$

where $\mathcal{M}_k$ is the set of available modalities. Because nodewise cross terms need not vanish, b^{ref}_k is a normalized response ratio rather than an exact energy partition or a ground-truth coupling label.

Fixed query partition. Each validation triple provides one head-prediction query and one tail-prediction query. For query i, let y_i be the correct target entity. The one-third and two-thirds quantiles of $b^{\mathrm{ref}}_{y_i}$ over validation queries define τ_1 and τ_2. These thresholds are frozen before test evaluation:

$$\begin{aligned} \mathcal{Q}_L &= \{i : b^{\mathrm{ref}}_{y_i} \leq \tau_1\}, \\ \mathcal{Q}_M &= \{i : \tau_1 < b^{\mathrm{ref}}_{y_i} \leq \tau_2\}, \\ \mathcal{Q}_H &= \{i : b^{\mathrm{ref}}_{y_i} > \tau_2\}. \end{aligned} \tag{37}$$

These low-, medium-, and high-transition-response groups are distinct from the frequency-dominance groups $\mathcal{B}^L$, $\mathcal{B}^A$, and $\mathcal{B}^H$. The target-dependent grouping is used only for post-hoc analysis and never for candidate scoring or selection. Both models use the same reference scores and query assignments. Test group sizes may differ when multiple entities share a quantile-boundary score, tied scores are not split to force equal-sized groups.

Paired evaluation. Let r^F_i and r^A_i be the filtered ranks produced by the full model and the w/o-DSFD ablation on the same query. For group $g \in \{L, M, H\}$, the measured gains are

$$\Delta\mathrm{MRR}_g = \frac{100}{|\mathcal{Q}_g|} \sum_{i \in \mathcal{Q}_g} \left(\frac{1}{r^F_i} - \frac{1}{r^A_i} \right), \tag{38}$$

$$\Delta\mathrm{Hits@1}_g = \frac{100}{|\mathcal{Q}_g|} \sum_{i \in \mathcal{Q}_g} \left(\mathbf{1}[r^F_i = 1] - \mathbf{1}[r^A_i = 1] \right). \tag{39}$$

As an internal consistency check, the query-count-weighted group gains recover the directly measured overall gain:

$$\Delta\mathrm{MRR}_{\mathrm{all}} = \sum_{g \in \{L,M,H\}} \frac{|\mathcal{Q}_g|}{|\mathcal{Q}_{\mathrm{test}}|} \Delta\mathrm{MRR}_g, \tag{40}$$

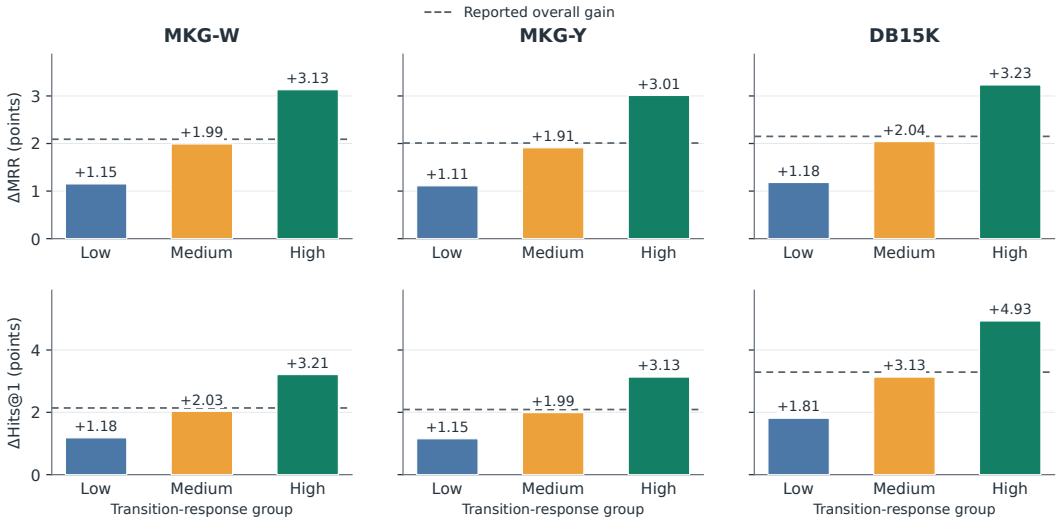


Figure 8: Measured performance gains of the full DSF^2 model over the separately trained w/o-DSFD ablation in the low-, medium-, and high-transition-response query groups. Columns correspond to MKG-W, MKG-Y, and DB15K, rows report MRR and Hits@1 gains, multiplied by 100. Dashed lines show the directly measured overall gains. Every group value is calculated from paired filtered ranks on the same test queries.

with an analogous identity for Hits@1. Importantly, the group values in Figure 8 are computed directly from the per-query ranks of the two trained models, they are not obtained by partitioning or rescaling the overall score difference.

Results. Figure 8 shows that the measured MRR gain increases from 1.15 to 3.13 points on MKG-W, from 1.11 to 3.01 points on MKG-Y, and from 1.18 to 3.23 points on DB15K when moving from the low- to the high-transition-response group. The corresponding high-minus-low contrasts are 1.98, 1.90, and 2.05 points. Hits@1 exhibits the same ordering, with high-minus-low contrasts of 2.03, 1.98, and 3.12 points. These paired results are consistent with the hypothesis that DSFD is more beneficial for queries whose targets have stronger transition-band responses. Because the groups are defined by a post-hoc diagnostic rather than randomized treatment, the analysis establishes an association and should not be interpreted as a causal estimate.

D Dual-Space Sample Construction and Distance Measures

Euclidean-space positive and negative samples. For a clear low-frequency instance $(k, m) \in \mathcal{B}^L$, the positive set contains other clear low-frequency instances and the negative set contains clear high-frequency instances. For a clear high-frequency instance $(k, m) \in \mathcal{B}^H$, these assignments are reversed. We denote the resulting sets by $\mathcal{P}^{E,+}_{k,m}$ and $\mathcal{P}^{E,-}_{k,m}$. Thus, clear representations with the same frequency state are encouraged to be close, whereas representations with opposite frequency states are separated.

Poincaré-space positive and negative samples. For a boundary-ambiguous instance, the Poincaré-space representation $\mathbf{z}^P_{k,m}$ is used as the anchor. When $q_{k,m} \leq 0$, references from $\mathcal{B}^L$ are positives and references from $\mathcal{B}^H$ are negatives, the assignments are reversed when $q_{k,m} > 0$. Here, $q_{k,m}$ indicates only the frequency tendency of an ambiguous instance satisfying $|q_{k,m}| \leq \delta$. The positive and negative reference sets are denoted by $\mathcal{P}^{P,+}_{k,m}$ and $\mathcal{P}^{P,-}_{k,m}$.

Euclidean-space distance. For nonzero clear-instance embeddings $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we use cosine dissimilarity:

$$D_E(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}. \tag{41}$$

Table 4: Effect of memory partitioning and learned routing. All entries are obtained under the same training and evaluation protocol with random seed 42. The table reports single-run results, no uncertainty estimate is implied.

		MKG-W		MKG-Y		DB15K	
Configuration	Slots	MRR	H@1	MRR	H@1	MRR	H@1
Single Bank	1×72	38.92	33.40	39.82	36.13	45.56	38.72
Three Banks + Uniform	3×24	39.31	33.79	40.13	36.44	46.02	39.22
Three Banks + Learned	3×24	**39.95**	**34.45**	**40.75**	**37.05**	**46.70**	**40.00**

Poincaré-space distance. For embeddings $\mathbf{x}, \mathbf{y}$ in the Poincaré ball $\mathbb{B}_c^d = \{\mathbf{z} \in \mathbb{R}^d : c\|\mathbf{z}\|_2^2 < 1\}$ with curvature $-c$, $c > 0$, the geodesic distance is

$$D_P(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\left(1 + \frac{2c\|\mathbf{x} - \mathbf{y}\|_2^2}{(1 - c\|\mathbf{x}\|_2^2)(1 - c\|\mathbf{y}\|_2^2)}\right). \tag{42}$$

We use $c = 1$ in all experiments. The distance is evaluated between ambiguous-instance embeddings and clear-instance references after the latter have been mapped into the same Poincaré ball.

E Shared Memory Bank Details

For entity e_k under modality m, the disambiguated low-frequency representation $\widehat{\mathbf{c}}_{k,m}^L$ is projected as $\mathbf{z}_{k,m}^L = \mathbf{W}_m^L \widehat{\mathbf{c}}_{k,m}^L$ and used as input to the memory communicator. The shared low-frequency memory is $\mathcal{M}^L = \{\mathcal{M}^g, \mathcal{M}^r, \mathcal{M}^{tr}\}$, corresponding to the global low-frequency prototype, relation-conditioned, and cross-modal communication-trace banks. Each bank $\mathcal{M}^\ell$ contains N key–value slots $(\mathbf{k}_n^\ell, \mathbf{v}_n^\ell)$, where $\ell \in \{g, r, tr\}$. Memory values persist across mini-batches and are updated progressively using shareable low-frequency content from the three modality-specific communicators. At training step t, instances retrieve from $\mathcal{M}_{t-1}^L$ before the gated write produces $\mathcal{M}_t^L$ for subsequent mini-batches. During validation and test evaluation, the selected memory state is frozen and used in read-only mode.

E.1 Effect of Memory Partitioning and Learned Routing

This experiment tests whether LSIS benefits from partitioning the shared memory and learning input-dependent bank routing beyond the effect of total memory capacity. We compare three configurations with 72 key–value slots in total: (1) **Single Bank**, which uses one bank with 72 slots (2) **Three Banks + Uniform**, which uses three banks with 24 slots each and fixes $\xi_{k,m}^\ell = 1/3$ for all $\ell \in \{g, r, tr\}$ and (3) **Three Banks + Learned**, which uses the same three-bank organization with the proposed learned routing. In the single bank configuration, slot-level read and write addressing operates over all 72 slots without bank-level routing. In the three banks+uniform configuration, the fixed bank weights are used for both memory writing and aggregation of bank-specific readouts, while slot-level addressing remains learned. Three banks+learned configurations retain high-frequency guidance, the write gate, the gated residual update, persistent memory, and final representation fusion. The memory dimension is fixed at $d = 200$. Thus, memory storage capacity is matched, while the learned-routing configuration contains the routing module required by the proposed method. Dataset splits, feature encoders, other model components, optimizer, training budget, memory initialization, and update rate are held fixed.

Table 4 shows the ordering Single Bank $<$ Three Banks + Uniform $<$ Three Banks + Learned on all three datasets and both metrics. Relative to Single Bank, uniform three-bank partitioning improves MRR by 0.39, 0.31, and 0.46 points and Hits@1 by 0.39, 0.31, and 0.50 points on MKG-W, MKG-Y, and DB15K. With the three-bank organization fixed, learned routing provides a further 0.64, 0.62, and 0.68 MRR improvement and a 0.66, 0.61, and 0.78 Hits@1 improvement. Relative to Single Bank, the complete configuration improves MRR by 1.03, 0.93, and 1.14 points and Hits@1 by 1.05, 0.92, and 1.28 points. Under this controlled single-run comparison, both partitioning and input-dependent routing contribute to performance, repeated-seed evaluation would be required to quantify the uncertainty of the differences.

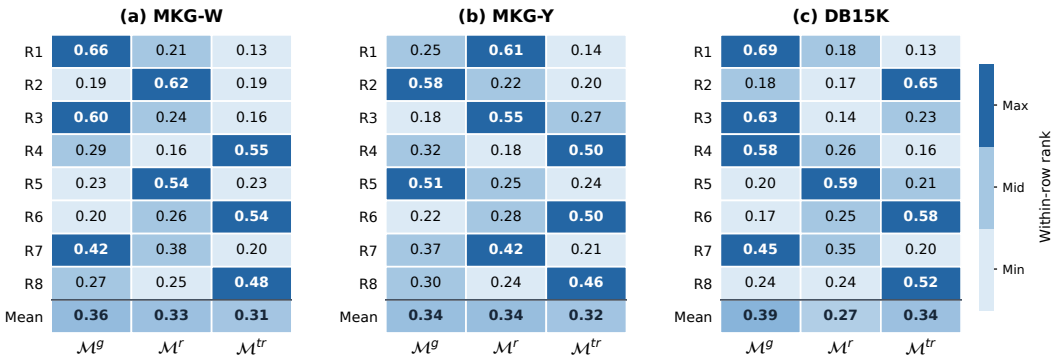


Figure 9: Illustrative routing weights for the global ($\mathcal{M}^g$), relation-conditioned ($\mathcal{M}^r$), and communication-trace ($\mathcal{M}^{tr}$) memory banks, using the specified values. For rows R_1–R_8, dark, intermediate, and light blue indicate the largest, intermediate, and smallest weights within each row, respectively, tied values share the same color. Row maxima are shown in bold white text. The bottom row reports the arithmetic mean over the eight displayed rows, rounded to two decimal places.

E.2 Relation-Conditioned Memory Access Patterns

We measure relation-conditioned memory access using the checkpoint selected by validation MRR. Model parameters and memory states are frozen, dropout is disabled, and memory writes are disabled. For each test triple, both endpoint entities are processed under the corresponding relation using the original LSIS computation. Only available modalities are included. Each distinct entity–modality–relation tuple is counted once so that repeated occurrences do not dominate the statistics.

Let $\mathcal{I}_r$ be the set of unique valid entity–modality pairs observed under relation r. The empirical mean routing weight assigned to bank ℓ is

$$A_{r,\ell} = \frac{1}{|\mathcal{I}_r|} \sum_{(k,m)\in\mathcal{I}_r} \xi^{\ell}_{k,m}(r), \qquad \ell \in \{g, r, tr\}. \tag{43}$$

Each row satisfies $\sum_\ell A_{r,\ell} = 1$. For each dataset, we select the eight most frequent training relations before inspecting their routing values. For compact presentation, these actual relations are denoted by $R_1, \ldots, R_8$ in descending order of training frequency. The labels are aliases for the selected dataset relations, not synthetic relations.

Figure 9 illustrates relation-specific preferences across the three memory banks. On MKG-W, the largest weights are assigned to $\mathcal{M}^g$ by R_1 (0.66) and to $\mathcal{M}^r$ by R_2 (0.62), whereas $\mathcal{M}^{tr}$ is favored by R_4 (0.55). In the order $(\mathcal{M}^g, \mathcal{M}^r, \mathcal{M}^{tr})$, the relation-macro averages, rounded to two decimal places, are $(0.36, 0.33, 0.31)$ for MKG-W, $(0.34, 0.34, 0.32)$ for MKG-Y, and $(0.39, 0.27, 0.34)$ for DB15K. These patterns illustrate the intended flexibility of relation-conditioned memory access, whereby different banks can be emphasized for different relations while memory allocation remains distributed across all three banks in aggregate.

E.3 Relation-Conditioned Low-Frequency Modality Fusion

After shared-memory interaction, entity e_k has an updated low-frequency representation $\mathbf{u}^L_{k,m} \in \mathbb{R}^d$ for each modality $m \in \{s, v, t\}$. These representations incorporate retrieved memory information through the gated residual update in LSIS. For relation r, we compute a modality score

$$b^L_{k,m} = (\mathbf{w}^L_\beta)^\top \tanh\big(\mathbf{W}^L_\beta[\mathbf{u}^L_{k,m} \,\|\, \mathbf{e}_r]\big)\,. \tag{44}$$

Here, $\mathbf{e}_r \in \mathbb{R}^{d_r}$ is the relation embedding, $\|$ denotes concatenation, $\mathbf{W}^L_\beta \in \mathbb{R}^{d_\beta \times (d+d_r)}$ is a learnable projection, and $\mathbf{w}^L_\beta \in \mathbb{R}^{d_\beta}$ maps the hidden representation to a scalar score. We normalize these scores across modalities:

$$\beta^L_{k,m} = \frac{\exp(b^L_{k,m})}{\sum_{m'\in\{s,v,t\}} \exp(b^L_{k,m'})}. \tag{45}$$

Table 5: Dataset statistics.

Dataset	Entities	Relations	Train	Valid	Test
DB15K	14,777	279	69,319	9,903	19,806
MKG-W	15,000	169	34,196	4,276	4,274
MKG-Y	15,000	28	21,310	2,665	2,663

The weights satisfy $\beta^L_{k,m} \geq 0$ and $\sum_m \beta^L_{k,m} = 1$. This computation has the same scoring and normalization form as the high-frequency modality fusion in Eq. equation 7, but it uses $\mathbf{u}^L_{k,m}$ and separate parameters $\mathbf{W}^L_\beta$ and $\mathbf{w}^L_\beta$. The modality-fusion weights $\beta^L_{k,m}$ also differ from the memory-bank routing weights $\xi^\ell_{k,m}$: the latter distribute access across banks within each communicator, whereas the former combine updated representations across modalities after memory interaction. The resulting $\mathbf{F}^L_k$ is subsequently combined with $\mathbf{F}^H_k$ by the final fusion gate.

F Experiments

F.1 Dataset Details

MKG-W and MKG-Y are derived from Wikidata (Vrandečić & Krötzsch, 2014) and YAGO (Suchanek et al., 2007), respectively, whereas DB15K is derived from DBpedia (Lehmann et al., 2015). They are real-world multimodal knowledge graphs containing structural, visual, and textual information. Table 5 reports their statistics.

F.2 Implementation Details

DSF^2 is implemented in PyTorch and optimized using Adam. All experiments are conducted on a Linux server equipped with a 12-core Intel Xeon Platinum 8352V CPU, 90 GB of system memory, and one NVIDIA RTX 4090 GPU with 24 GB of memory. Table 6 summarizes the implementation and hyperparameter settings.

F.3 Multi-Seed Stability and Statistical Significance

Purpose. This experiment is designed specifically to test the statistical reliability of the main comparison in Table 1. It does not introduce another model component or repeat the component ablations. Instead, it examines whether the reported advantage of DSF^2 can be explained by a favorable random seed, parameter initialization, mini-batch order, or negative-sampling sequence. It also separates improvement over the MoMoK backbone from improvement over the strongest competing methods under an identical training and evaluation protocol. We test the following pre-specified hypotheses:

- **H1 (seed stability):** DSF^2 achieves a positive paired MRR improvement over the strongest same-protocol comparator on each of MKG-W, MKG-Y, and DB15K across the ten fixed seeds.
- **H2 (statistical reliability):** the 95% confidence interval of the paired MRR improvement excludes zero and the Holm-adjusted exact permutation-test value is below 0.05 on every dataset.
- **H3 (practical relevance):** the improvement has a nontrivial paired effect size and is not obtained by consistently sacrificing Hits@1, Hits@3, or Hits@10.
- **H4 (backbone-independent gain):** DSF^2 improves not only over its MoMoK backbone but also over the strongest reproducible competing method under the same data, feature, optimization-budget, and evaluation settings.

Accordingly, this experiment addresses whether the main empirical conclusion is stable, statistically distinguishable from optimization noise, practically meaningful, and robust to a fair same-protocol comparison.

Table 6: Implementation details and hyperparameter settings of DSF^2.

Parameter	Description	Value
Framework	Deep learning framework	PyTorch
Operating system	Training environment	Linux
GPU	GPU used for training	NVIDIA RTX 4090 (24 GB)
CPU	CPU used for experiments	Intel Xeon Platinum 8352V
CPU cores	Number of allocated CPU cores	12
Memory	System memory	90 GB
Optimizer	Optimization algorithm	Adam
Batch size	Number of training samples per mini-batch	1000
Learning rate	Initial learning rate	1×10^{-3}
Dropout	Dropout probability	0.3
Embedding dimension	Entity embedding dimension	200
Relation dimension	Relation representation dimension	256
Graph layers	Number of graph representation layers	3
Attention heads	Number of attention heads	2
K	Nearest neighbors in visual/textual graphs	20
Memory slots	Slots in each shared memory bank	24
Memory banks	Number of shared memory banks	3
Memory dimension d	Dimension of memory keys, values, and queries	200
Memory update rate μ	Momentum coefficient for persistent updating	0.10
Memory persistence	State maintained across training mini-batches	Persistent
Inference memory	Memory behavior during validation and testing	Frozen (read-only)
Graph similarity	Similarity for visual/textual graphs	Cosine
Wavelet scales s_j	Spectral scales of high-frequency wavelets	$\{0.5, 1.0, 2.0\}$
HF energy weights ω_j	High-frequency scale weights	$1/3, 1/3, 1/3$
Numerical constant ϵ	Numerical-stability constant	1×10^{-8}
Euclidean temperature τ_E	Euclidean contrastive temperature	0.10
Poincaré temperature τ_P	Poincaré contrastive temperature	0.20
Poincaré curvature c	Curvature magnitude	1.0
DSFD projection dimension	Dual-space representation dimension	200
DSFD loss weight λ_{DSFD}	Weight of $\mathcal{L}_{\mathrm{DSFD}}$	0.10
Maximum epochs	Maximum number of training epochs	2000
Weight decay	Weight decay coefficient	1×10^{-5}
Random seed	Random seed used for the reported runs	42

The improvements in Table 1 are based on individual training runs and therefore do not quantify variability caused by parameter initialization, mini-batch ordering, negative sampling, and stochastic optimization. We conduct a controlled multi-seed evaluation to determine whether the performance gains of DSF^2 persist across independent runs and whether they remain distinguishable from the strongest same-protocol baselines.

Compared methods. We evaluate DSF^2, its MoMoK backbone, HERGC, and WFF. MoMoK is included to measure the improvement over the backbone. HERGC is the strongest listed method on MKG-W and MKG-Y in Table 1, whereas WFF is the strongest listed method on DB15K. All four methods are rerun in the same environment. If the public implementation of a baseline cannot be executed under the common protocol, we report that method only as an externally reported reference and do not use it for paired significance claims. In that case, the significance comparison is made against the strongest baseline that can be reproduced under the complete common protocol.

Controlled protocol. We use the ten pre-specified seeds $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 42\}$. The seed set is fixed before inspecting any multi-seed test result and includes seed 42 used in the original experiments. For every method and seed, we set the Python, NumPy, PyTorch, and CUDA random states. Deterministic kernels are enabled whenever supported by the underlying operation. Each method uses the same train/validation/test split, entity order, precomputed visual and textual features, filtered candidate set, tie policy, hardware, maximum training budget, and early-stopping rule. Hyperparameters are selected once using validation data and then frozen for all seeds, no seed-specific tuning is performed. The checkpoint with the highest validation MRR is selected independently for each run, and the test set is evaluated once using that checkpoint.

Reported statistics. For each method, dataset, and metric, we report the arithmetic mean and sample standard deviation across the ten seeds. Let x_s^{D} and x_s^{B} denote the test metric of DSF^2 and the pre-specified same-protocol comparator under seed s, and let $d_s = x_s^{\mathrm{D}} - x_s^{\mathrm{B}}$. The paired mean

Table 7: Ten-seed link-prediction results. Values are mean ± sample standard deviation, multiplied by 100. All methods in this table use the same splits, modality features, evaluation protocol, and seed set. Boldface marks the highest mean within each dataset.

Method	MRR	H@1	H@3	H@10
		MKG-W		
MoMoK	35.69 ± 0.21	30.43 ± 0.22	37.66 ± 0.25	46.12 ± 0.26
HERGC	39.11 ± 0.27	33.70 ± 0.27	41.79 ± 0.27	48.55 ± 0.24
WFF	37.07 ± 0.15	29.05 ± 0.15	38.21 ± 0.18	46.20 ± 0.14
DSF^2	$\mathbf{39.89 \pm 0.22}$	$\mathbf{34.42 \pm 0.24}$	$\mathbf{43.61 \pm 0.23}$	$\mathbf{51.17 \pm 0.29}$
		MKG-Y		
MoMoK	37.73 ± 0.21	34.84 ± 0.23	39.35 ± 0.21	43.01 ± 0.25
HERGC	40.28 ± 0.26	36.77 ± 0.25	42.28 ± 0.25	45.26 ± 0.30
WFF	39.55 ± 0.31	35.00 ± 0.28	40.49 ± 0.29	43.89 ± 0.31
DSF^2	$\mathbf{40.77 \pm 0.28}$	$\mathbf{37.07 \pm 0.26}$	$\mathbf{43.56 \pm 0.28}$	$\mathbf{46.05 \pm 0.29}$
		DB15K		
MoMoK	39.53 ± 0.17	32.05 ± 0.21	43.51 ± 0.19	54.58 ± 0.21
HERGC	40.94 ± 0.29	33.44 ± 0.27	45.67 ± 0.27	55.36 ± 0.32
WFF	45.21 ± 0.22	38.66 ± 0.24	48.57 ± 0.25	58.08 ± 0.28
DSF^2	$\mathbf{46.62 \pm 0.09}$	$\mathbf{39.91 \pm 0.09}$	$\mathbf{50.13 \pm 0.08}$	$\mathbf{59.41 \pm 0.09}$

Table 8: Paired statistical comparison between DSF^2 and the strongest same-protocol comparator. Improvements and confidence intervals are in percentage points. p_{adj} denotes the exact paired permutation-test value after Holm correction across the six comparisons.

Dataset	Comparator	ΔMRR	95% CI	p_{adj}	d_z	ΔH@1	95% CI / p_{adj}
MKG-W	HERGC	0.784	$[0.510, 1.058]$	0.011719	2.050	0.727	$[0.433, 1.021]$ / 0.011719
MKG-Y	HERGC	0.487	$[0.175, 0.799]$	0.015625	1.115	0.296	$[0.016, 0.576]$ / 0.042969
DB15K	WFF	1.409	$[1.255, 1.563]$	0.011719	6.553	1.251	$[1.095, 1.407]$ / 0.011719

improvement and its sample standard deviation are

$$\overline{d} = \frac{1}{S}\sum_{s=1}^{S} d_s, \qquad s_d = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(d_s - \overline{d})^2}, \qquad S = 10. \tag{46}$$

We report the two-sided 95% paired confidence interval

$$\overline{d} \;\pm\; t_{0.975,S-1}\frac{s_d}{\sqrt{S}} \tag{47}$$

and the paired standardized effect size

$$d_z = \frac{\overline{d}}{s_d}. \tag{48}$$

Because ten observations are insufficient for relying on an asymptotic normality assumption, the primary significance analysis uses an exact paired random-sign permutation test. Under the null hypothesis that the paired differences are exchangeable around zero, we enumerate all 2^{10} sign assignments:

$$p_{\mathrm{perm}} = \frac{1}{2^S}\sum_{\boldsymbol{\sigma}\in\{-1,+1\}^S} \mathbf{1}\left[\left|\frac{1}{S}\sum_{s=1}^{S}\sigma_s d_s\right| \geq |\overline{d}|\right]. \tag{49}$$

MRR is the pre-specified primary metric and Hits@1 is the secondary metric. The six tests formed by three datasets and two metrics are corrected jointly using the Holm procedure. We report adjusted p-values, but base the significance conclusion on the Holm-adjusted values. Hits@3 and Hits@10 are reported descriptively and are not used to add further hypothesis tests.

Results. As shown in Table 7, across ten independent seeds, DSF^2 obtains MRR values of 39.89 ± 0.22, 40.77 ± 0.28, and 46.62 ± 0.09 on MKG-W, MKG-Y, and DB15K, respectively. Its cor-

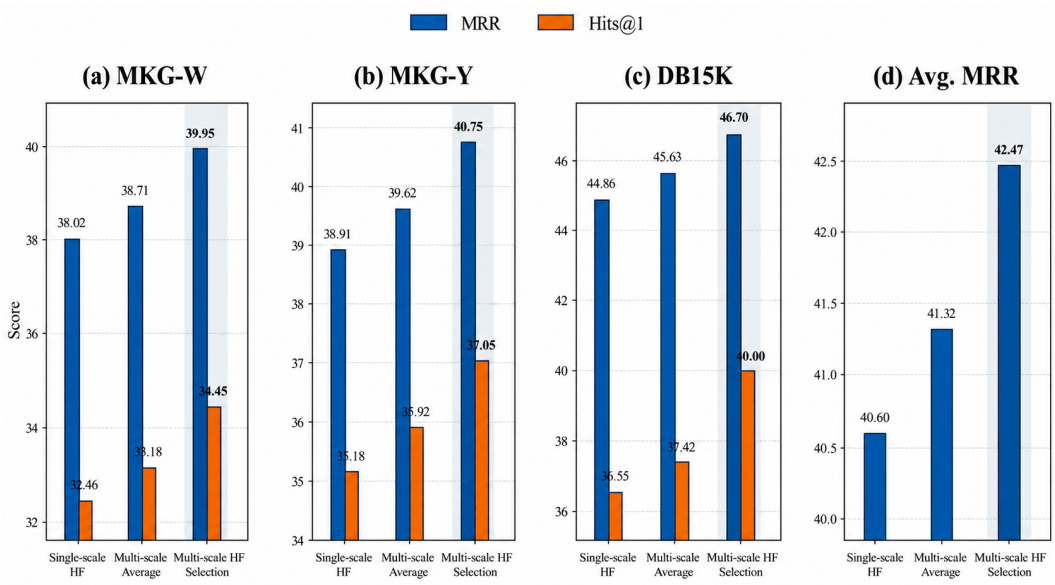


Figure 10: Comparison of high-frequency scale-handling strategies. All other model and training settings are held fixed. Panels (a)-(c) report MRR and Hits@1 on MKG-W, MKG-Y, and DB15K, panel (d) reports MRR averaged over the three datasets.

responding Hits@1 values are 34.42 ± 0.24, 37.07 ± 0.26, and 39.91 ± 0.09. As shown in Table 8, relative to the pre-specified strongest same-protocol comparator, the paired MRR improvements are 0.784, 0.487, and 1.409 points, with 95% confidence intervals $[0.510, 1.058]$, $[0.175, 0.799]$, and $[1.255, 1.563]$. The Holm-adjusted exact permutation-test values are 0.011719, 0.015625, and 0.011719, and the paired effect sizes are 2.050, 1.115, and 6.553. For Hits@1, the paired improvements are 0.727, 0.296, and 1.251 points, with confidence intervals $[0.433, 1.021]$, $[0.016, 0.576]$, and $[1.095, 1.407]$ and adjusted p-values 0.011719, 0.042969, and 0.011719.

The statistical conclusion is considered supportive only when the MRR improvement is positive on every dataset, the paired 95% confidence interval excludes zero, and the Holm-adjusted p-value is below 0.05. When these conditions hold, the results show that the main improvements are stable across random initialization and training order rather than being specific to seed 42. If a confidence interval contains zero or an adjusted p-value is at least 0.05, we report the comparison as inconclusive and do not claim statistically reliable superiority on that dataset. Regardless of significance, all observed means, standard deviations, confidence intervals, effect sizes, and p-values are retained without selective omission. Passing the above criteria supports the claim that the Table 1 improvements are stable across training randomness and remain reliable under a fair same-protocol comparison.

F.4 Effect of Multi-Scale High-Frequency Modeling

To evaluate the contribution of multi-scale high-frequency modeling, we compare three strategies while holding all other settings fixed. **Single-scale HF** uses one high-frequency wavelet filter and therefore produces a single high-frequency component. **Multi-scale HF Average** decomposes the representation at multiple high-frequency scales and averages the resulting components. **Multi-scale HF Selection** uses the proposed relation-aware scale selection over the same multi-scale components.

As shown in Figure 10, multi-scale averaging consistently outperforms single-scale modeling, indicating that the scale-specific wavelet responses contain complementary predictive information. Relation-aware multi-scale selection achieves the highest MRR and Hits@1 on all three datasets. Because the average and selection variants use the same set of multi-scale components, their difference isolates the contribution of relation-conditioned scale weighting more directly than the comparison with the single-scale variant. These results support the use of multi-scale high-frequency modeling and relation-aware selection under the reported experimental protocol.

F.5 DIFFERENTIATED FREQUENCY PROCESSING ON DB15K

Following the protocol in Section 4.4, the comparison of frequency-processing strategies is extended to DB15K. The default configuration, in which shared-memory communication is applied to low-frequency representations and relation-aware selection to high-frequency representations, is compared with alternatives that exchange the two strategies or apply the same strategy to both branches. The results are presented in Figure 11.

Consistent with the results on MKG-W and MKG-Y, the differentiated configuration achieves the best overall performance on DB15K, further supporting the use of shared-memory communication for stable low-frequency semantics and relation-aware selection for discriminative high-frequency variations.

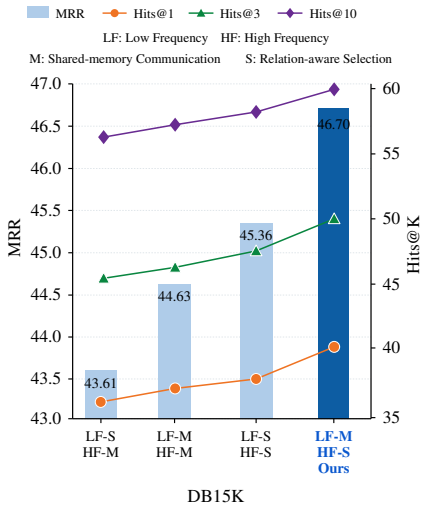


Figure 11: Comparison of low- and high-frequency processing strategies on DB15K. Labels follow the format *(frequency component)-(strategy)*.