Beyond Alignment: Discrepancy-driven Consistency Reasoning for Commonsense Question Answering

Anonymous ACL submission

Abstract

Leveraging pre-trained language models (LMs) together with knowledge graphs (KGs) has become a common paradigm for commonsense question answering (CommonsenseQA), as it combines linguistic understanding with structured factual knowledge. In this setting, the core challenge lies not in the availability of external knowledge, but in selecting the candidate answer that best satisfies both linguistic semantic constraints and structured knowledge constraints. Existing language model–knowledge graph (LM–KG) approaches predominantly rely on cross-modal alignment, assuming that increased representational agreement leads to improved reasoning. However, in CommonsenseQA, semantic discrepancies between language models and knowledge graphs frequently indicate constraint violations, where linguistic plausibility conflicts with graph structure or vice versa. By smoothing out such discrepancies, alignment-centric methods obscure critical reasoning signals. To address this limitation, we propose Semantic-Aware Reasoning Network (DCRN), which explicitly models semantic discrepancy between LMs and KGs as a reasoning signal rather than noise. DCRN leverages discrepancy-aware learning to support more reliable joint reasoning for Question Answering.

1 Introduction

Using pre-trained Language Models (LMs) together with Knowledge Graphs (KGs) has become one of the mainstream paradigms for commonsense question answering (CommonsenseQA). LMs provide strong contextual understanding and broad parametric knowledge, while KGs offer structured factual relations and explicit connectivity. Combining them is expected to support more reliable commonsense reasoning, especially when questions require multi-hop inference or factual grounding.

Despite this progress, joint reasoning over LMs and KGs remains challenging. Prior work often aggregates encoded representations from both modalities and computes similarity scores to guide answer selection (Fang et al., 2020; Yasunaga et al., 2021). Such aggregation can work when evidence from the two modalities is consistent, but it is less effective for reasoning-intensive commonsense questions, where multiple candidates can be linguistically plausible and only subtle constraints distinguish the correct answer.

To improve joint reasoning, recent studies have explored stronger cross-modal interaction mechanisms and alignment. JointLK (Sun et al., 2022) employs a bidirectional attention module to facilitate information exchange, while GreaseLM (Zhang et al., 2021) uses shared interaction layers for fusion. QAT (Park et al., 2023) models structural and semantic relations via meta-paths and applies self-attention for reasoning, and GRT (Zhao et al., 2024) further incorporates knowledge triples as graph features to reduce modality gaps. Although these methods improve cross-modal interaction, they largely follow an alignment-centric paradigm, which emphasizes increasing representational agreement between LMs and KGs.

However, the core challenge in CommonsenseQA is not simply to align two modalities. It is to select the candidate answer that is most consistent with both linguistic semantics and structured knowledge. These discrepancies arise because the two modalities encode complementary but different signals: LMs capture fine-grained contextual semantics but may lack explicit factual grounding, whereas KGs provide reliable relational knowledge but have limited capacity for nuanced semantic interpretation. In many cases, such discrepancies indicate constraint violations, where an answer is plausible under one modality but conflicts with the other. Alignment-centric approaches that smooth out these discrepancies can remove useful cues for distinguishing correct answers from hard distractors.

Motivated by this, we propose a discrepancy-driven reasoning framework, **DCRN** (**D**iscrepancy-driven **C**onsistency **R**easoning **N**etwork), for CommonsenseQA. DCRN addresses the inherent semantic differences between language models (LMs) and knowledge graphs (KGs) by explicitly exposing and utilizing cross-modal discrepancies for complementary reasoning, rather than enforcing representational alignment. Specifically, we feed contextual representations from LMs and semantic representations from KGs into a Semantic Discrepancy-Aware Reconstruction module, which leverages cross-modal semantic conditioning and reconstruction-based supervision to isolate informative discrepancy residuals. Furthermore, these residual semantic discrepancies are transformed into explicit representations through an Semantic Fusion Mapping, enabling discrepancy-aware consistency reasoning for answer selection.

Our **contributions** are summarized as follows:

- We propose **DCRN**, a discrepancy-driven consistency reasoning framework for CommonsenseQA that explicitly models cross-modal semantic discrepancies between language models and knowledge graphs.
- We introduce Semantic Discrepancy-Aware Reconstruction, which exposes and structures semantic residuals induced by cross-modal conditioning.
- We propose an Semantic Fusion Mapping module that transforms residual discrepancies into first-class semantic features for consistency-based reasoning.
- Extensive experiments on four cross-domain QA benchmarks show that DCRN achieves state-of-the-art performance on commonsense and medical question answering tasks, demonstrating the effectiveness of discrepancy-driven consistency reasoning.

2 Preliminaries

2.1 LM+KG Question Answering

Language model and knowledge graph based question answering (LM+KG QA) combines pre-trained language models (LMs) and structured knowledge graphs (KGs) for reasoning-intensive tasks. Given a question q and candidate answers $C = a_1, \ldots, a_n$, the LM encodes the question–answer text to capture contextual semantics, while a task-specific subgraph $g = (V, E)$ is extracted from the KG and encoded to model relational knowledge. The textual and graph representations are then integrated to estimate the plausibility of each answer choice, typically through alignment, fusion, or joint reasoning mechanisms, forming a general framework for commonsense and knowledge-intensive question answering.

2.2 Cross-modal Reasoning Paradigms

Cross-modal reasoning in LM+KG QA aims to leverage both contextual semantics from language models and structured relational knowledge from graphs. In this work, we adopt a discrepancy-driven reasoning paradigm that does not enforce early alignment between modalities. Instead, we allow LM and KG representations to interact under mutual semantic constraints, expose their residual discrepancies through reconstruction, and explicitly represent such discrepancies for downstream reasoning. This paradigm enables the model to reason over cross-modal consistency rather than relying solely on aligned representations.

3 Methodlogy

3.1 Semantic Discrepancy-Aware Reconstruction

We introduce a discrepancy-aware semantic reconstruction mechanism with two components. **Cross-Modal Semantic Conditioning** allows LM and KG representations to interact and expose semantic discrepancies, while the **Learnable Routing Filter** selectively reconstructs representations by preserving informative discrepancy signals and suppressing redundant cross-modal information.

Cross-Modal Semantic Conditioning. To expose semantic discrepancies between language models (LMs) and knowledge graphs (KGs), DCRN performs cross-modal semantic conditioning that allows each modality to interpret the input under the semantic constraints of the other, without enforcing early alignment. This process preserves modality-specific inductive biases while making cross-modal semantic residuals observable.

Given an input context sequence $\{w_i\}_{i=1}^{M}$, we obtain contextual token representations using a pre-trained language model:

$$\{\widetilde{t}_0, \widetilde{t}_1, \ldots, \widetilde{t}_M\} = \mathbf{Encoder}_{\mathrm{LM}}(\{w_1, \ldots, w_M\}), \quad (1)$$

where $\{\widetilde{t}_i\}_{i=1}^{M} \in \mathbb{R}^D$ denotes contextual token embeddings and $\widetilde{t}_0$ corresponds to the *[CLS]* token.

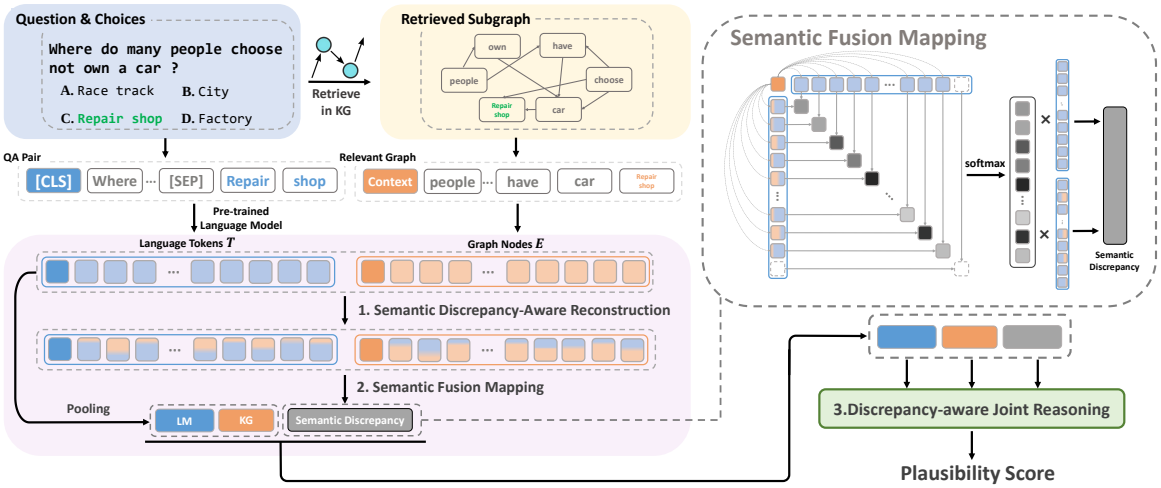

Figure 1: Overall architecture of DRCN. The framework processes question-answer pairs and retrieved subgraphs through three main stages: (1) Semantic Discrepancy-Aware Reconstruction captures semantic gaps between textual tokens and graph nodes, (2) Semantic Fusion Mapping integrates discrepancy signals to enhance representations, (3) Discrepancy-aware Joint Reasoning combines enhanced representations to generate the final plausibility score.

In parallel, entity nodes in the retrieved knowledge subgraph are encoded as $\mathbf{E} = \{e_i\}_{i=1}^{|V|} \in \mathbb{R}^D$ following (Fang et al., 2020). The central subgraph node is initialized with the *[CLS]* representation to establish a shared semantic anchor across modalities.

To capture fine-grained intra-modal semantics, we apply text and graph semantic refinement at each layer using Text Semantic Conditioning (TSC) and Graph Semantic Conditioning (GSC) modules:

$$\{\widetilde{t}_1^l, \ldots, \widetilde{t}_M^l\} = \mathbf{TSC}(\{t_1^{l-1}, \ldots, t_M^{l-1}\}), \quad (2)$$

$$\{\widetilde{e}_1^l, \ldots, \widetilde{e}_{|V|}^l\} = \mathbf{GSC}(\{e_1^{l-1}, \ldots, e_{|V|}^{l-1}\}), \quad (3)$$

where $\{t_i^{l-1}\}$ and $\{e_i^{l-1}\}$ denote text and graph representations from the previous layer.

To enable cross-modal semantic conditioning, we employ an asymmetric attention mechanism that conditions one modality on the semantic space of the other. Unlike symmetric fusion or self-attention, this design allows modality-specific semantics to be selectively interpreted under cross-modal constraints, making unresolved semantic differences explicit. The conditioning operation is defined as:

$$\boldsymbol{Q} = \boldsymbol{X}W_{\boldsymbol{Q}}, \quad \boldsymbol{K} = \tilde{\boldsymbol{X}}W_{\boldsymbol{K}}, \quad \boldsymbol{V} = \tilde{\boldsymbol{X}}W_{\boldsymbol{V}}, \quad (4)$$

$$\hat{\mathcal{Z}} = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{V}, \quad (5)$$

where $W_{\boldsymbol{Q}}$, $W_{\boldsymbol{K}}$, and $W_{\boldsymbol{V}}$ are learnable projection matrices and d is the key dimension. Here, $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are alternately instantiated as text representations $\mathbf{T}_{\text{LM}}$ or graph representations $\mathbf{E}_{\text{KG}}$, yielding cross-modal conditioned representations $\hat{\mathcal{Z}}$ that explicitly encode discrepancy-aware semantic information.

Learnable Routing Filter. Following cross-modal semantic conditioning, the injected representations contain both complementary semantic signals and unresolved cross-modal discrepancies. To selectively reconstruct informative semantics while preserving discrepancy-aware residuals, DCRN introduces a learnable routing filter that adaptively controls how cross-modal information is incorporated into each modality.

Specifically, we apply gated transformations to the conditioned representations to disentangle recoverable semantic components from residual discrepancies:

$$\widetilde{t}_i = f_{\text{down}}\left(\sigma\left(f_{\text{gate}}(t_i)\right) \odot f_{\text{up}}(t_i)\right), \quad (6)$$

$$\widetilde{e}_i = f_{\text{down}}\left(\sigma\left(f_{\text{gate}}(e_i)\right) \odot f_{\text{up}}(e_i)\right), \quad (7)$$

where $f_{\text{gate}}(\cdot)$, $f_{\text{up}}(\cdot)$, and $f_{\text{down}}(\cdot)$ denote learnable linear transformations parameterized by $\mathbf{W}_g$, $\mathbf{W}_u$, and $\mathbf{W}_d$, respectively. The gating function $\sigma(\cdot)$ determines the extent to which injected cross-modal information contributes to semantic reconstruction, and $\odot$ denotes element-wise multiplication.

The reconstructed representations $\widetilde{\mathbf{T}} = \{\widetilde{t}_i\}_{i=1}^{M}$ and $\widetilde{\mathbf{E}} = \{\widetilde{e}_i\}_{i=1}^{|V|}$ are then routed back to their original semantic spaces through learnable combi-

nation:

$$\overset{*}{\mathbf{T}} = \alpha_1 \cdot \mathbf{T} + \alpha_2 \cdot \widetilde{\mathbf{T}}, \quad \overset{*}{\mathbf{E}} = \beta_1 \cdot \mathbf{E} + \beta_2 \cdot \widetilde{\mathbf{E}}, \tag{8}$$

where α_1, α_2, β_1, and β_2 are learnable scalar coefficients that regulate the balance between original modality-specific semantics and discrepancy-aware reconstructed signals. Through this routing mechanism, DCRN selectively preserves semantic residuals induced by cross-modal conditioning while recovering complementary information, enabling fine-grained discrepancy-aware reasoning across modalities.

3.2 Semantic Fusion Mapping

After N layers of discrepancy-aware semantic reconstruction, the resulting representations encode cross-modal information with implicit discrepancy cues. To integrate such information for downstream reasoning without suppressing informative semantic differences, DCRN introduces a *Semantic Fusion Mapping* module that selectively fuses original and reconstructed representations under discrepancy-aware constraints, consisting of Masked Semantic Mapping and Semantic Discrepancy Completion.

Masked Semantic Mapping. To control the fusion process and mitigate interference from irrelevant semantic units, we construct bidirectional visibility matrices to guide LM-to-KG and KG-to-LM interactions. Tokens and graph nodes are treated as semantic units, and asymmetric masking strategies are applied to regulate which units participate in cross-modal fusion.

In the source semantic space, we mask all units except semantically enriched intermediate nodes to focus fusion on informative representations. In the target semantic space, we mask the central unit while retaining peripheral semantic units, preventing trivial self-alignment while preserving contextual diversity. Each entry $M_{i,j}$ in the visibility matrix M is set to 1 for visible pairs and 0 otherwise, and transformed into an attention mask:

$$\widetilde{M}_{i,j} = \begin{cases} 0 & \text{if } M_{i,j} = 1 \\ -\infty & \text{if } M_{i,j} = 0 \end{cases} \tag{9}$$

Based on $\widetilde{M}$, masked attention is applied to selectively fuse cross-modal semantic information while preserving discrepancy-relevant signals:

$$\alpha = \mathrm{softmax}\left(\frac{(h^i W_q)(h^j W_k)^\top}{\sqrt{d}} + \widetilde{M} \right), \tag{10}$$

where W_q and W_k are projection matrices, d denotes the key dimension, and $\widetilde{M}$ governs cross-modal visibility during fusion.

Semantic Discrepancy Completion. Using the discrepancy-aware attention weights, we further model the semantic differences between original and reconstructed representations to capture complementary cross-modal information introduced during fusion. The semantic discrepancy vector is computed as:

$$\widetilde{D} = f_{\mathrm{d}} \left(\alpha \cdot T \,;\, \alpha \cdot \widetilde{T} \right), \tag{11}$$

where T and $\widetilde{T}$ denote the original and reconstructed textual representations, respectively, α represents the attention weights derived from semantic fusion mapping, and f_{d} is a learnable transformation that integrates the fused representations.

3.3 Consistency Reasoning with Semantic Discrepancy Supervision

With explicit discrepancy representations, DCRN performs consistency-based reasoning to select the final answer by jointly considering textual, structural, and discrepancy-aware signals. A dedicated training objective further regulates discrepancy behavior, encouraging semantic consistency for aligned cases while preserving discriminative differences when semantic conflicts arise.

Semantic Discrepancy Contrastive Loss.

To maintain semantic consistency while enhancing discriminative capabilities, we introduce a contrastive loss based on mean squared error that measures element-wise differences between reconstructed and original features. We apply differential supervision for positive and negative samples to enhance semantic contrast. For positive samples (correctly matched LM-KG pairs), the model minimizes reconstruction discrepancy to preserve semantic consistency. Conversely, for negative samples (mismatched pairs), the model retains semantic differences to prevent over-alignment and maintain cross-modal discriminability. This contrastive supervision guides the model to perform meaningful reconstruction for aligned inputs while preserving discriminative power across modalities.

$$\mathcal{L}_{\mathrm{SDC}} = \frac{1}{m} \sum_{i=1}^{m} \left\| V_i' - \tilde{V}_i^0 \right\|_2^2 + \lambda \cdot \frac{1}{n} \sum_{i=1}^{n} \left\| T_i' - \tilde{T}_i^0 \right\|_2^2 \tag{12}$$

Here, V_i' and $\tilde{V}_i^0$ denote the reconstructed and original KG features for the i-th positive sample, while

T_i' and $\tilde{T}_i^0$ represent the reconstructed and original LM features for the i-th negative sample. The coefficient λ controls the relative importance of the negative loss component. This design guides the model to restore semantic content for aligned pairs while retaining discriminative discrepancies for misaligned inputs.

Discrepancy-aware Joint Reasoning. Finally, discrepancy-aware semantic representations are integrated with textual and graph representations for answer selection. The answer probability is computed as:

$$p = P(a \mid q) = \mathrm{MLP}([s;\ g;\ t]), \qquad (13)$$

where s, g, and t denote semantic discrepancy, graph, and textual representations, respectively. The output probabilities are normalized across candidate answers using softmax.

The overall training objective combines answer supervision with discrepancy-aware consistency supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \beta \mathcal{L}_{\mathrm{SDC}}, \qquad (14)$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss and β controls the influence of semantic discrepancy supervision.

Methods	IHdev-Acc.(%)	IHtest-Acc.(%)
RoBERTa-large(w/o KG)	73.07(±0.45)	68.69(±0.56)
+ RGCN	72.69(±0.19)	68.41(±0.66)
+ MHGRN	74.45(±0.10)	71.11(±0.81)
+ QA-GNN	76.54(±0.21)	73.41(±0.92)
+ GreaseLM	78.50(±0.50)	74.20(±0.40)
+ JointLK	77.88(±0.25)	74.43(±0.83)
+ GSC	79.11(±0.22)	74.48(±0.41)
+ SEPTA	79.61(±0.17)	74.78(±0.23)
+ QAT	79.50(±0.40)	75.40(±0.30)
+ GRT	79.60(±0.30)	76.10(±0.40)
+ DCRN(Ours)	**80.29(±0.20)**	**76.48(±0.35)**

Table 1: Performance comparison on *CommonsenseQA* in-house split. We follow the data division method of (Lin et al., 2019) and report the in-house Dev(IHdev) and Test (IHtest) accuracy(mean and standard deviation of four runs).

4 Experiment

4.1 Datasets

We evaluate our method on four question-answering datasets: CommonsenseQA(Talmor et al., 2019),OpenBookQA(Mihaylov et al., 2018), Riddle(Lin et al., 2021) and MedQA-USMLE(Jin et al., 2021).Dataset statistics are in the supplement.

CommonsenseQA. This dataset contains questions that require commonsense reasoning. Since the offcial test set labels are not publicly available, we mainly report performance on the in-house development (IHdev) and test (IHtest) sets following (Lin et al., 2019). For CommonsenseQA,we adopt ConceptNet (Speer et al., 2017) as the structured knowledge source.

OpenBookQA. This dataset requires reasoning with elementary science knowledge. We experiment on the offcial data split from (Mihaylov et al., 2018).we adopt ConceptNet as the structured knowledge source.

RiddleSense. This dataset requires complex commonsense reasoning abilities, an understanding of figurative language and counterfactual reasoning skills.We experiment on the offcial data split from (Lin et al., 2021).we adopt ConceptNet as the structured knowledge source.

MedQA-USMLE. This dataset originates from the United States Medical License Exam (USMLE) practice sets, requiring biomedical and clinical knowledge. Thus, we utilize a knowledge graph provided by (Yasunaga et al., 2021). We use the same data split as(Jin et al., 2021).

4.2 Baseline Methods

We focus on approaches that enhance model architecture and LM-KG reasoning interfaces. We compare DCRN against the following methods, all employing the same pre-trained LM as text encoder: (1) RGCN (Schlichtkrull et al., 2018), (2) MHGRN (Fang et al., 2020), (3) QAGNN (Yasunaga et al., 2021), (4) GSC (Wang et al., 2022), (5) JointLK (Sun et al., 2022), (6) GreaseLM (Zhang et al., 2021), (7) QAT (Park et al., 2023), (8) GRT (Zhao et al., 2024), (9) SEPTA (Peng et al., 2024). Methods (1)-(4) perform joint reasoning using graph and textual representations, while methods (5)-(6) incorporate explicit cross-modal information exchange. Methods (7)-(8) enhance reasoning by modeling edge features, and method (9) introduces subgraph retrieval strategies for cross-modal integration.

4.3 Main Results

We present experimental results comparing DCRN with baseline methods on CommonsenseQA, OpenbookQA, RiddleSense, and MedQA-USMLE datasets.

General Domain. On CommonsenseQA (Table **??**), DCRN establishes new state-of-the-art

Methods	RoBERTa-Large	AristoRoBERTa
LMs (w/o KG)	64.80 (±2.37)	78.40 (±1.64)
+ RGCN	62.45 (±1.57)	74.60 (±2.53)
+ MHGRN	66.85 (±1.19)	80.60 (±NA)
+ QA-GNN	67.80 (±2.75)	82.77 (±1.56)
+ GreaseLM	-	84.80 (±NA)
+ GSC	70.33 (±0.81)	86.67(±0.46)
+ JointLK	70.34 (±0.75)	84.92 (±1.07)
+ QAT	71.20 (±0.80)	86.90 (±0.20)
+ SEPTA	72.33 (±0.35)	87.37 (±0.51)
+ GRT	72.60 (±1.00)	87.30 (±0.80)
+ DCRN(Ours)	**72.80(±0.60)**	**87.60(±0.54)**

Table 2: Test accuracy comparison on OpenBookQA.

Methods	Accuracy
RoBERTa-Large(w/o KG)	52.6
MHGRN	54.5
QA-GNN	67.0
GreaseLM	67.2
JointLK	67.3
SEPTA	67.6
DCRN(Ours)	**70.6**

Table 3: Test accuracy comparison on *RiddleSense*.QAT and GRT do not provide preprocessed data for *RiddleSense*. Therefore we can't train the QAT and GRT model on *RiddleSense*.

Methods	Accuracy
SapBERT-Base (w/o KG)	37.2
QA-GNN	38.0
GreaseLM	38.5
QAT	39.3
GRT	39.5
DCRN (ours)	**40.8**

Table 4: Test accuracy comparison on *MedQA-USMLE*.

performance, surpassing fine-tuned LMs by 7.7% on IH-test accuracy. Notably, DCRN outperforms strong baselines including GRT, QAT, and SEPTA, demonstrating the effectiveness of semantic discrepancy-aware reasoning.

Table **??** shows results on OpenBookQA using the RoBERTa-large backbone. DCRN achieves a 0.5% absolute improvement over previous methods. When using the larger AristoRoBERTa model with additional training data, DCRN maintains a 0.3% gain, demonstrating robustness across different language models.

DR	SFM	DL	IHtest-Acc.(%)
✓	✓	✓	**76.48 (±0.35)**
✓	✓		75.82 (±0.40)
✓			72.40 (±0.50)
			69.85 (±0.64)

Table 5: Ablation study on *CommonsenseQA*.

Question Types	QAT	GRT	DCRN
Full question set	79.5	79.6	**80.3 (↑0.7)**
Question w/ negation	79.0	79.7	**80.0 (↑0.3)**
Question w/ entities ≤ 7	79.9	79.1	**80.0 (↑0.1)**
Question w/ entities > 7	79.7	80.0	**80.6 (↑0.6)**

Table 6: Accuracy of datasets for questions involving complex reasoning such as negation terms, more entity mentions. DCRN consistently outperforms the KG-augmented QA models (QAT and GRT) in these complex reasoning settings.

On the semantically complex RiddleSense dataset (Table 3), DCRN outperforms the previous best model (SEPTA) by 3.0%, highlighting its enhanced capacity for handling intricate semantic reasoning. These results demonstrate DCRN's superior adaptability across diverse datasets and language models, particularly on complex reasoning tasks.

Biomedical Domain. To evaluate cross-domain generalizability, we test DCRN on MedQA-USMLE. As shown in Table 4, DCRN achieves 40.8% accuracy, representing a 1.3% improvement over the strongest baseline (GRT). These results confirm DCRN's effectiveness in capturing domain-specific semantics across both textual and graph representations.

5 Analysis

In this section, we provide an in-depth analysis of **DCRN** to answer the following research questions: (1) How does each component in DCRN contribute to discrepancy-driven consistency reasoning? (2) Does explicitly modeling semantic discrepancy improve cross-modal commonsense reasoning? (3) In what types of reasoning scenarios does semantic discrepancy provide the greatest benefit?

5.1 Ablation Studies

To address **Question (1)**, we conduct ablation studies on DCRN by progressively removing key

components that correspond to different stages of discrepancy-driven reasoning. Specifically, removing the **Semantic Discrepancy Contrastive Loss**(DL) eliminates consistency-aware supervision over discrepancy behavior. Ablating the **Semantic Fusion Mapping**(SFM) module prevents the construction of explicit discrepancy representations, forcing the model to rely solely on aligned semantic features. Removing the **Discrepancy Reconstruction**(DR) module disables cross-modal semantic conditioning and residual structuring entirely.

The results are reported in Table 5. When all discrepancy-related components are removed, performance drops by 6.6%, indicating that discrepancy modeling plays a central role in DCRN. Among individual components, removing the Semantic Mapping module leads to the largest performance degradation (3.9%), highlighting the importance of explicitly representing semantic residuals rather than implicitly absorbing them through fusion. These findings suggest that discrepancy-aware representations are critical for enabling complementary reasoning across language and knowledge modalities.

5.2 Effectiveness of Discrepancy Modeling

To answer **Question (2)**, we evaluate whether explicitly modeling semantic discrepancy improves cross-modal commonsense reasoning. We compare **DCRN** with representative knowledge-enhanced baselines, including QAT and GRT, and further analyze performance under varying levels of semantic complexity. In particular, we group questions by the number of involved entities and the presence of negation, both of which introduce competing or conflicting semantic constraints.

As shown in Table 6, DCRN consistently outperforms baseline methods across all evaluated categories, achieving accuracies of 80.0%, 80.0%, and 80.6%. The performance gains are most pronounced for questions containing negation or involving more than seven entities, where simple alignment between language and knowledge representations is insufficient. These results indicate that semantic discrepancy modeling is especially effective in scenarios where language cues and structured knowledge impose partially inconsistent constraints, supporting the motivation that discrepancy serves as a critical signal for consistency-based reasoning.

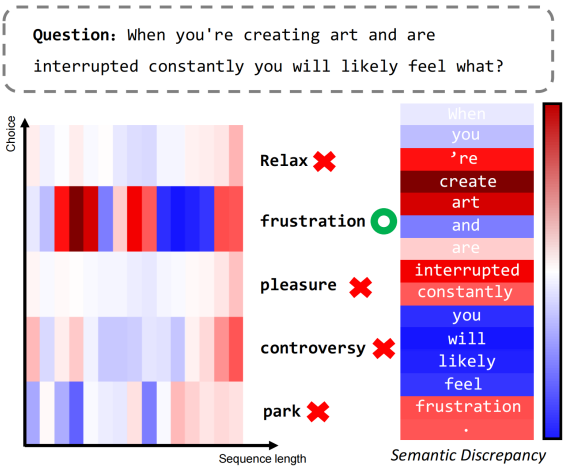

Figure 2: How semantic discrepancy enhances cross-modal reasoning. We visualize the enhancement changes for each question–answer pair, focusing on the selected answer choice.

5.3 Qualitative Analysis

To address **Question (3)**, we conduct a qualitative analysis to examine how semantic discrepancies are activated and utilized during joint reasoning. For each question–answer pair, DCRN retrieves a corresponding knowledge subgraph and constructs explicit semantic discrepancy representations between the language model and the knowledge graph. We visualize attention weights produced by the **Semantic Mapping** module, where red indicates semantic enhancement and blue denotes suppression.

We first observe that discrepancy activation patterns differ markedly between correct and incorrect answer choices. As illustrated in Figure 2, for the correct answer *"frustration"*, DCRN assigns high discrepancy-aware attention to tokens such as *"create"*, *"art"*, and *"interrupted"*. These tokens reflect semantic tensions between the linguistic context and the retrieved knowledge, signaling unresolved constraints that are informative for reasoning. At the same time, attention to irrelevant or weakly related tokens is effectively suppressed.

In contrast, for incorrect answer candidates, discrepancy signals remain weak and diffuse. This indicates that DCRN does not indiscriminately amplify cross-modal differences, but instead selectively activates discrepancy representations when semantic conflicts are meaningful for inference. As a result, incorrect question–choice pairs do not exhibit strong enhancement or suppression patterns.

To further analyze how discrepancy influences

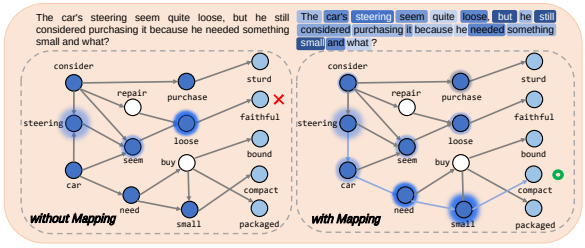

Figure 3: A case of semantic discrepancy enhanced reasoning. The question and corresponding answer are: "The car's steering seemed quite loose, but he still considered purchasing it because he needed something small and what?" and "compact". For simplicity, we present only a subset of entities in the figure and omit synonym-substituted entities.

reasoning trajectories, Figure 3 visualizes node importance weights from the pooling module, where the halo size around each node reflects its contribution to inference. Without semantic discrepancy modeling, the model tends to over-attend to semantically peripheral nodes such as *"seem"* or *"loose"*, which may divert reasoning toward loosely associated concepts. In contrast, when discrepancy representations are incorporated, attention shifts toward nodes and tokens that participate in semantic conflict resolution. Co-occurring markers such as *"but"*, *"needed"*, and *"small"* interact to activate discrepancy-aware semantics, leading to more coherent reasoning paths.

Overall, these qualitative results demonstrate that discrepancy-driven enhancement does not introduce spurious conflicts. Instead, it selectively amplifies semantic signals that bridge gaps between the linguistic context and structured knowledge. By reasoning over such discrepancies, DCRN effectively resolves competing constraints, resulting in more reliable cross-modal inference.

6 Related Work

6.1 Question Answering with LM+KG

Knowledge graphs (KGs) encode structured relational knowledge and are commonly combined with language models (LMs) for knowledge-intensive question answering. Prior work (Yasunaga et al., 2021; Fang et al., 2020) performs joint reasoning by integrating KG representations modeled with graph neural networks (GNNs) and textual representations from LMs. To strengthen cross-modal interaction, subsequent studies introduce explicit interaction layers (Sun et al., 2022; Zhang et al., 2021) or adopt unified self-attention mechanisms with alternative KG encodings (Park et al., 2023; Zhao et al., 2024). Across these approaches, a common design choice is to align representations from LMs and KGs to support joint reasoning.Recent studies take initial steps toward modeling and exploiting such differences for complementary reasoning in commonsense question answering (Wang et al., 2023).

6.2 Cross-modal Discrepancy Modeling

Information from heterogeneous modalities often exhibits discrepancies at both representational and semantic levels, reflecting differences in inductive biases and information organization. If not properly handled, such discrepancies may lead models to over-rely on a single modality and limit effective cross-modal reasoning. Prior work has therefore explored explicitly modeling semantic differences or residual signals to support complementary understanding across modalities (Xu et al., 2020; Deng et al., 2024; Yang et al., 2024; Baltrusaitis et al., 2019). These studies, primarily in multi-modal settings involving vision, audio, and text, show that representing cross-modal discrepancies can help recover missing information and improve joint inference.Cross-modal discrepancy in LM–KG reasoning reflects differences in semantic abstraction and structural awareness across modalities (Dai et al., 2024; Mavromatis et al., 2024).

7 Conclusion

This work revisits joint reasoning with language models and knowledge graphs for commonsense question answering, arguing that effective reasoning requires going beyond cross-modal alignment. We observe that semantic discrepancies between language and graph representations are common and informative, and propose DCRN to explicitly model such discrepancies for consistency-based reasoning. Extensive experiments show that leveraging semantic discrepancy leads to consistent improvements across multiple benchmarks.

Limitations

This work focuses on leveraging semantic discrepancies between language models and knowledge graphs to support consistency-based reasoning for commonsense question answering. While DCRN demonstrates consistent improvements across multiple benchmarks, it relies on the availability and quality of external knowledge graphs. In scenarios

where the underlying knowledge graph is sparse, noisy, or misaligned with the question domain, the effectiveness of discrepancy modeling may be limited.In addition, DCRN explicitly models cross-modal semantic discrepancies at the representation level, which may not capture all forms of higher-order reasoning discrepancies, such as those involving implicit commonsense assumptions or long-range causal dependencies that are not well represented in either modality.Finally, the proposed framework introduces additional computational overhead due to cross-modal reconstruction and discrepancy-aware learning components. Although this overhead is moderate in our experiments, further optimization would be necessary for deployment in large-scale or latency-sensitive settings.

Ethics Statement

This work focuses on developing methods for improving commonsense question answering by jointly reasoning over language models and knowledge graphs. All experiments are conducted on publicly available benchmark datasets, and no new data involving human subjects is collected. The proposed approach does not introduce additional risks related to privacy, security, or personal data misuse. Potential biases present in the results may originate from the underlying language models or knowledge graphs used in the experiments, which reflect the biases of their training data. While our method aims to improve reasoning reliability by modeling semantic discrepancies, it does not explicitly address fairness or bias mitigation.

References

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *NAR*, 32(suppl_1):D267–D270.

Peter Clark, Oren Etzioni, Tushar Khot, Daniel Khashabi, Bhavana Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, et al. 2020. From 'f' to 'a' on the N.Y. regents science exams: An overview of the aristo project. *AI Magazine*, 41(4):39–53.

Ruiting Dai, Yuqiao Tan, Lisi Mo, Shuang Liang, Guohao Huo, Jiayi Luo, and Yao Cheng. 2024. G-SAP: graph-based structure-aware prompt learning over heterogeneous knowledge for commonsense reasoning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR 2024, Phuket, Thailand, June 10-14, 2024*, pages 1051–1060. ACM.

Jinhong Deng, Xiaoyue Zhang, Wen Li, Lixin Duan, and Dong Xu. 2024. Cross-domain detection transformer based on spatial-aware and semantic-aware token alignment. *IEEE Transactions on Multimedia*, 26:5234–5245.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP-IJCNLP*.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *NAACL-HLT*.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Costas Mavromatis, Petros Karypis, and George Karypis. 2024. Sempool: Simple, robust, and interpretable KG pooling for enhancing language models. In *Advances in Knowledge Discovery and Data Mining - 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2024, Taipei, Taiwan, May 7-10, 2024, Proceedings, Part IV*, volume 14648 of *Lecture Notes in Computer Science*, pages 154–166. Springer.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP*.

Jinyoung Park, Hyeong Kyu Choi, Juyeon Ko, Hyeonjin Park, Ji-Hoon Kim, Jisu Jeong, Kyungmin Kim, and Hyunwoo Kim. 2023. Relation-aware language-graph transformer for question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13457–13464.

Boci Peng, Yongchao Liu, Xiaohe Bo, Sheng Tian, Baokun Wang, Chuntao Hong, and Yan Zhang. 2024. Subgraph retrieval enhanced by graph-text alignment for commonsense question answering. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 39–56, Cham. Springer Nature Switzerland.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2022. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5049–5060, Seattle, United States. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Kuan Wang, Yuyu Zhang, Diyi Yang, Le Song, and Tao Qin. 2022. GNN is a counter? revisiting GNN for question answering. In *ICLR*.

Yujie Wang, Hu Zhang, Jiye Liang, and Ru Li. 2023. Dynamic heterogeneous-graph reasoning with language models and knowledge representation learning for commonsense question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14048–14063. Association for Computational Linguistics.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *NAR*, 46(Database-Issue):D1074–D1082.

Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020. Cross-modal attention with semantic consistence for image–text matching. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5412–5425.

Yuchen Yang, Yu Wang, and Yanfeng Wang. 2024. SDA: Semantic discrepancy alignment for text-conditioned image retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5250–5261, Bangkok, Thailand. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2021. Greaselm: Graph reasoning enhanced language models. In *International Conference on Learning Representations*.

Ruilin Zhao, Feng Zhao, Liang Hu, and Guandong Xu. 2024. Graph reasoning transformers for knowledge-aware question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19652–19660.

Dataset	# Questions	# Choices per Question
CommonsenseQA	12,102	5
OpenBookQA	5,957	4
Riddle-Sense	5,715	5
MedQA-USMLE	12,723	4

Table 7: Dataset Statistics.

A Dataset Statistics

We provide details on the datasets and knowledge graph we have adopted for our experiments. In Table 7, we summarized the question and answer choice statistics for the three datasets we experimented on. For CommonsenseQA (Talmor et al., 2019) and OpenBookQA (Mihaylov et al., 2018) datasets, we used the ConceptNet (Speer et al., 2017) as the knowledge graph. This is a general-domain knowledge graph retaining 799,273 nodes and 2,487,810 edges in total. Each edge is assigned to a relation type, which is merged as in Table 8 following (Fang et al., 2020; Yasunaga et al., 2021). On the other hand, MedQA-USMLE (Bodenreider, 2004) requires external biomedical knowledge. We thus used a different knowledge graph provided by (Wishart et al., 2018), which contains 9,958 nodes and 44,561 edges. Given each QA context, we extract a subgraph from the full knowledge graph following (Yasunaga et al., 2021).

B Experimental Settings

In this section, we provide details for preprocess of KG and LMs, as well as hyperparameters we used in our experiments.

B.1 KG Retrieval

Given each input text segment W, we follow the procedure from (Yasunaga et al., 2021) to retrieve a relevant local KG G from the raw KG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. First, we use the entity linker from the spaCy library to link entity mentions in W to entity nodes in $\mathcal{G}$, obtaining an initial set of nodes V_{el}. Second, we add any bridge entities in $\mathcal{G}$ that are in a 2-hop path between any pair of linked entities in V_{el} to get the total retrieved nodes $V \subseteq \mathcal{V}$. If the number of nodes in V exceeds 200, we prune V by randomly sampling 200 nodes from it to be the final retrieved nodes V. Lastly, we retrieve all the edges in $\mathcal{G}$ that connect any two nodes in V to obtain $E \subseteq \mathcal{E}$, forming the final local KG, $G = (V, E)$.

Relation	Merged Relation
AtLocation LocatedNear	AtLocation
Causes CausesDesire *MotivatedByGoal	Causes
Antonym DistinctFrom	Antonym
HasSubevent HasFirstSubevent HasLastSubevent HasPrerequisite Entails MannerOf	HasSubevent
IsA InstanceOf DefinedAs	IsA
PartOf *HasA	PartOf
RelatedTo SimilarTo Synonym	RelatedTo

Table 8: Merged Relations. *RelationX indicates the reverse relation of RelationX.

B.2 Graph Initialization

For the ConceptNet knowledge graph used in the general commonsense domain, we follow the method of MHGRN (Fang et al., 2020) to prepare the initial KG node embeddings. Specifically, we convert triplets in the KG into sentences using predefined templates for each relation. Then, these sentences are fed into BERT-Large to compute embeddings for each sentence. Finally, for each entity, we collect all sentences containing the entity, extract all token representations of the entity's mention spans in these sentences, and return the mean pooling of these representations.

For the UMLS knowledge graph used in the biomedical domain, node embeddings are initialized similarly using the pooled token output embeddings of the entity name from SapBERT (Liu et al., 2021).

B.3 Language Models

For CommonsenseQA, OpenBookQA and Riddle-Sense, we take advantage of the pretrained

RoBERTa-large (Liu et al., 2019) model. In the case of OpenBookQA, we additionally apply AristoRoBERTa (Clark et al., 2020) which utilizes textual data as an external source of information. For the MedQA-USMLE dataset, SapBERT (Liu et al., 2021)is used in place of RoBERTa models.

B.4 Implementation Details

In our experiments, the hyperparameters are tuned with respect to the development set in each dataset, and evaluated on the test set. For training, one GeForce RTX 4090 is used. We used RAdam as our optimizer using a linear learning rate scheduler with a warmup phase on four datasets. Following prior works, we took the performance mean and standard deviation with different seed in 0,1,2,3,4,5,6 We kept identical hyperparameter settings across seeds and their settings vary by dataset, which are specified in Table **??** and Table 11

B.5 Implementation

Building on prior work (Yasunaga et al., 2021), we employ ConceptNet (Speer et al., 2017) as our structured knowledge source for the three general-domain QA tasks. Following the preprocessing pipeline of (Fang et al., 2020), we retrieve subgraphs from ConceptNet with up to three hops per question.The subgraph of MedQA is constructed via integrating the Disease Database portion of the Unified Medical Language System (Bodenreider, 2004) and the DrugBank (Wishart et al., 2018) knowledge graph. Training of DCRN uses the RAdam optimizer (Liu et al., 2020). To ensure a fair comparison, all baselines and our DCRN variants share the same underlying language model. We tune the LM's learning rate over $\{1\times10^{-5}, 3\times10^{-5}, 4\times10^{-5}\}$ and DCRN's learning rate over $\{1\times10^{-3}, 3\times10^{-4}\}$. Each model is trained on a single NVIDIA RTX 4090 GPU with a batch size of 128, requiring on average 5 hours to converge.

B.6 Complex Questions

To validate the effectiveness of the question answering models on complex questions, we experimented with diverse question types such as questions with negation, questions with fewer entities (≤ 7), and questions with more entities (> 7) following (Sun et al., 2022) in Table 5 (below) of the main paper. We selected questions with negation terms by retrieving questions that contain (no, not, nothing, never, unlikely, don't, doesn't, didn't, can't, couldn't) from the CommonsenseQA IHdev set.

Hyperparameter	CSQA	OBQA
epochs	22	100
freeze LM epochs	2	2
tolerance epochs	8	30
warmup steps	100	150
batch size	128	128
Discrepancy Reconstruction layers	5	5
KG-conditioned attention heads	4	4
LM-conditioned attention heads	8	8
Cross-modal conditioning heads	8	8
hidden dimension	200	200
embedding dropout	0.2	0.2
reconstruction dropout	0.2	0.2
MLP dropout	0.2	0.2
learning rate	1e-3	1e-3
discrepancy loss weight (λ)	50	50
routing coefficient (α)	0.1	0.1
LM learning rate	1e-5	3e-5
weight decay	1e-2	1e-2
gradient norm clip	1	1
learning rate schedule	warm-up linear decay	warm-up linear decay

Table 9: Hyperparameter Settings for CSQA and OBQA.

Hyperparameter	RiddleSense
epochs	36
freeze LM epochs	2
tolerance epochs	10
warmup steps	100
batch size	128
Discrepancy Reconstruction layers	3
KG-conditioned attention heads	4
LM-conditioned attention heads	8
Cross-modal conditioning heads	8
hidden dimension	200
embedding dropout	0.2
reconstruction dropout	0.2
MLP dropout	0.2
learning rate	3e-4
discrepancy loss weight (λ)	50
routing coefficient (α)	0.1
LM learning rate	3e-5
weight decay	1e-2
gradient norm clip	1
learning rate schedule	warm-up linear decay

Table 10: Hyperparameter Settings for RiddleSense.

C Additional Experimental Results

In this section, we provide additional analyses to show the robustness of our SARN on diverse hyperparameter settings including loss balance coefficient, λ.

Hyperparameter	MedQA-USMLE
epochs	34
freeze LM epochs	0
tolerance epochs	10
warmup steps	500
batch size	128
Discrepancy Reconstruction layers	3
KG-conditioned attention heads	4
LM-conditioned attention heads	8
Cross-modal conditioning heads	8
hidden dimension	200
embedding dropout	0.2
reconstruction dropout	0.2
MLP dropout	0.2
learning rate	4e-5
discrepancy loss weight (λ)	50
routing coefficient (α)	0.1
LM learning rate	5e-5
weight decay	1e-2
gradient norm clip	1
learning rate schedule	warm-up linear decay

Table 11: Hyperparameter Settings for the Biomedical Domain.

Methods	Acc.
RoBERTa	72.1
ALBERT	76.5
RoBERTa + FreeLB	73.1
RoBERTa + HyKAS	73.2
RoBERTa + KE	73.3
RoBERTa + KEDGN	74.4
XLNet + GraphReason	75.3
RoBERTa + MHGRN	75.4
ALBERT + PG	75.6
RoBERTa + QA-GNN	76.1
RoBERTa + JointLK	76.6
RoBERTa + GSC	76.4
RoBERTa + DCRN **(ours)**	**78.2**

Table 12: Performance comparison on *CommonsenseQA* official leaderboard.

C.1 Leaderboard Evaluation Results

We further report SARN's performance on the official leaderboard for two datasets, CommonsenseQA and OpenBookQA. In Table 12, In Table 13

C.2 Ablation Studies

To evaluate the robustness of our model under different hyperparameter configurations, we conducted a sensitivity analysis on key hyperparameters. Specifically, we varied the value of λ to examine its effect on performance on CommonsenseQA. Recall that λ is the weighting coefficient between the Semantic Discrepancy Contrast Loss and the Cross-Entropy Loss. As shown in Table 7, the best Dev accuracy is achieved when λ = 50. We also analyzed the impact of the number of layers used in Semantic Discrepancy Reconstruction and found that the best Dev accuracy is obtained when M = 5. Moreover, we observe that even under extreme hyperparameter settings, the performance does not degrade drastically.

Methods	Acc.
ALBERT	81.0
AristoRoBERTa	77.8
HGN	81.4
AMR-SG	81.6
ALBERT + KPG	81.8
AristoRoBERTa + QA-GNN	82.8
T5	83.2
T5 + KB	85.4
UnifiedQA	87.2
GreaseLM	84.8
AristoRoBERTa + JointLK	85.6
AristoRoBERTa + GSC	87.4
AristoRoBERTa + QAT	87.6
AristoRoBERTa + DCRN **(ours)**	**88.0**

Table 13: *OpenBookQA* official learderboard. We also report the official leaderboard performance on the test dataset.

Ablation Type	Ablation	Dev Acc.
Number of DR layers (M)	$M = 4$	79.6 ($\pm$0.4)
	$M = 5$	**80.3 ($\pm$0.2)**
	$M = 6$	79.4 ($\pm$0.3)
	$M = 7$	78.8 ($\pm$0.5)
Loss coefficient (λ)	$\lambda = 0.1$	79.5 ($\pm$0.5)
	$\lambda = 1$	79.4 ($\pm$0.5)
	$\lambda = 10$	79.9 ($\pm$0.3)
	$\lambda = 50$	**80.3 ($\pm$0.2)**
	$\lambda = 100$	79.1 ($\pm$0.4)

Table 14: **Ablation study** of our model Hyperparameter, using the CommonsenseQA IH-dev set.