Benchmarking Diverse-Modal Entity Linking with Generative Models

Sijia Wang[1*], **Alexander Hanbo Li**[2†], **Henry Zhu**[2], **Sheng Zhang**[2], **Chung-Wei Hang**[2], **Pramuditha Perera**[2], **Jie Ma**[2], **William Wang**[2], **Zhiguo Wang**[2], **Vittorio Castelli**[2]
Bing Xiang[2], **Patrick Ng**[2]
[1] Virginia Tech [2] AWS AI Labs
sijiawang@vt.edu
{hanboli,henghui,zshe,cwhang,pramudi,jieman,wyw, zhiguow,vittorca,bxiang,patricng}@amazon.com

Abstract

Entities can be expressed in diverse formats, such as texts, images, or column names and cell values in tables. While existing entity linking (**EL**) models work well on per modality configuration, such as text-only EL, visual grounding, or schema linking, it is more challenging to design a unified model for diverse modality configurations. To bring various modality configurations together, we constructed a benchmark for diverse-modal EL (**DMEL**) from existing EL datasets, covering all three modalities including text, image, and table. To approach the DMEL task, we proposed a generative diverse-modal model (**GDMM**) following a multimodal-encoder-decoder paradigm. Pre-training GDMM with rich corpora builds a solid foundation for DMEL without storing the entire KB for inference. Fine-tuning GDMM builds a stronger DMEL baseline, outperforming state-of-the-art task-specific EL models by 8.51 F1 score on average. Additionally, extensive error analyses are conducted to highlight the challenges of DMEL, facilitating future research on this task.

1 Introduction

Linking ambiguous mentions to unambiguous referent in a knowledge base (KB) such as Wikipedia, known as **Entity linking** (EL) (Shen et al., 2015), is an essential component for applications like question answering (Ferrucci, 2012; Chen et al., 2017; Lewis et al., 2020) and recommendation systems (Yang et al., 2018). **Diverse-Modal Entity Linking** (DMEL) extends the scope of interest from textual entity linking to heterogeneous input formats, such as linking visual and textual expressions to KB (Adjali et al., 2020b,a; Moon et al., 2018; Gan et al., 2021a; Zheng et al., 2022; Wang et al., 2022d; Gan et al., 2021a; Cui et al., 2021) and linking mentions in natural language to tables or database (DB) schema (Liu et al., 2021; Katsakioris et al., 2022; Shi et al., 2020; Lei et al., 2020; Chen et al., 2020; Wang et al., 2022a). Figure 1 demonstrate three examples of DMEL, including (a) classical textual entity linking, (b) textual-visual entity linking in which the question or mentions are paired with image(s), and (c) tabular schema linking in which the mentions are linked to column names or cell values.

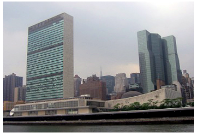
<table>
<tr><th></th><th>Information</th><th></th><th>DMEL Annotations</th></tr>
<tr><td>(a)</td><td colspan="2">Two years later, Ireland competed for the first time in a World Cup Finals, at Italia '90</td><td><b>Italia '90:</b> 1990 FIFA World Cup</td></tr>
<tr><td>(b)</td><td></td><td>UN Headquarters in New York</td><td><b>UN</b>: United Nations
<b>New York</b>: New York City</td></tr>
<tr><td>(c)</td><td><table><tr><th>Coach</th><th>Years</th><th>Wins</th><th>Losses</th></tr><tr><td>Ed Sabre</td><td>1920</td><td>3</td><td>1</td></tr><tr><td>Mack Erwin</td><td>1957–60</td><td>62</td><td>30</td></tr><tr><td>Fred Jordan</td><td>1992-</td><td>726</td><td>552</td></tr></table></td><td>What was the number of loses while Fred Jordan was coaching?</td><td><b>loses</b> ➤ Losses
<b>coaching</b> ➤ Coach</td></tr>
</table>

Figure 1: DMEL examples for (a) textual EL, (b) textual-visual EL, and (c) tabular schema linking.

Retrieval-based contrastive learning or ranking mechanism is the mainstream for early visual entity linking by leveraging a matching score between the mention and the KB entities (Cui et al., 2021; Wang et al., 2022d; Zheng et al., 2022). However, these methods require storage of dense representations of all KB entities, and when the size of entities increases (e.g. Wikipedia has 6M articles), it raises concerns for space complexity and also inference-time latency. Meanwhile, linking mentions to tables or DB schemes, known as schema linking, remains an important but under-explored task. For example, in text-to-SQL generations, incorrect schema linking usually counts for a large portion of the errors (Zhong et al., 2017; Yu et al., 2018; Shi et al., 2020; Lei et al., 2020; Taniguchi et al., 2021). Previous string matching heuristic

[*] Work conducted during an internship at Amazon.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 7841–7857
July 9-14, 2023 ©2023 Association for Computational Linguistics

(Chen et al., 2020) or embedding matching methods (Chen et al., 2020; Wang et al., 2022a; Guo et al., 2019; Wang et al., 2020) lack semantic and schema understanding, and can hardly generalize well to new domains. Last but not the least, previous endeavors of entity linking are limited to individual tasks including textual EL, textual-visual EL, or schema linking, and lack a general view for the DMEL problem.

To this end, we propose a unified DMEL task that includes existing EL datasets on all three modalities – text, image, and table. The unified DMEL task is challenging because the model needs to handle a wide spectrum of modality configurations together. On the modeling side, because storing all entity information (e.g. all the images in the entire KB) is expensive at inference time, we propose to use a unified generative model that can take diverse-modal input and generate entity names in an autoregressive fashion. Additionally, the mention diversity and ambiguity issue in schema linking can be addressed by pre-training the generative model.

In this work, we build a generic diverse-modal architecture for end-to-end DMEL. The DMEL dataset is constructed from five existing datasets, including GERBIL benchmark, WikiDiverse, MELBench-Wikipedia, Squall, and SLSQL, covering diverse EL tasks. The proposed generative diverse-modal model (**GDMM**) is first pre-trained on large-scale text corpus BLINK and images corpus from Wikipedia KB, offering profound prior knowledge. Extensive experiments are then conducted on the DMEL benchmark to compare our proposed generative model to previous state-of-the-art methods. Experimental results show that GDMM achieves strong performance on the DMEL dataset and outperforms state-of-the-art task-specific EL models. Our contributions include:

- We define a novel diverse-modal Entity Linking task, which links an entity mention within heterogeneous information sources to a knowledge base. A unified dataset is constructed for rigorous DMEL examination.
- A generative diverse-modal model GDMM is proposed following a multimodal-encoder-decoder structure. The multimodal encoder allows collective representation between each modality. The autoregressive structure enables us to directly predict the entity name without storing the entire KB. The pre-training experimental results confirm that a candidate trie created from entity names is sufficient for inference.
- The experimental results show that the proposed model obtains state-of-the-art performance on (almost) each individual EL task.

Dataset	**Modality**	**Size**		
		#L	**#V**	**#U**
GERBIL	L → L	42,854	0	0
WikiDiverse	LV → L	7,823	6,924	0
MELBench	LV → L	18,880	18,880	0
Squall	LU → L	11,274	0	2,108
SLSQL	LU → L	8,034	0	166
DMEL	LVU → L	88,865	25,804	2,274

Table 1: Comparison of existing datasets. Statistics for modality, (L)anguage, (V)ision, and Tabl(U)r, are shown.

2 Problem Formulation

We assume to have a KB (e.g., Wikipedia or a DB schema) where each entity is a unique entry in the KB. We formulate the following DMEL problem: given a multimodal input $\{\mathbf{x}_i, \mathbf{v}_i, \mathbf{u}_i\}$ of textual (L), visual (V), and tabular (U) modality respectively, an entity mention $\mathbf{m}_i$ within the input, and a candidate set $\mathcal{C}_i = \{\mathbf{c}_i^1, \cdots, \mathbf{c}_i^K\}$, the task is to link the mention $\mathbf{m}_i$ to one entity in $\mathcal{C}_i$. We assume the entity span is given. Sometimes the candidate set can be the entire entity collection $\mathcal{E}$. Particular instances of DMEL problem include but are not limited to: *Textual Entity Disambiguation* where a given mention $\mathbf{m}_i$ in $\mathbf{x}_i$ will be linked to one entity in $\mathcal{C}_i$; *Textual-Visual Entity Disambiguation* where the $\mathbf{m}_i$ and a given image $\mathbf{v}_i$ will be linked to one entity in $\mathcal{C}_i$; *Schema linking* where a $\mathbf{m}_i$ in a SQL query will be linked to table schema, i.e., a column name within given tables $\mathbf{u}_i$. If the mention is not a valid entity or not in $\mathcal{C}_i$, the target label is “nil”.

3 DMEL Benchmark

We build the DMEL benchmark from five existing datasets, including GERBIL benchmark (Verborgh et al., 2018), WikiDiverse (Wang et al., 2022d), MELBench-Wikidata (as MELBench in the rest of the paper) (Gan et al., 2021a), Squall (Shi et al., 2020), and SLSQL (Lei et al., 2020). We evaluate textual-visual entity disambiguation capability on WikiDiverse and MELBench, and evaluate tabular

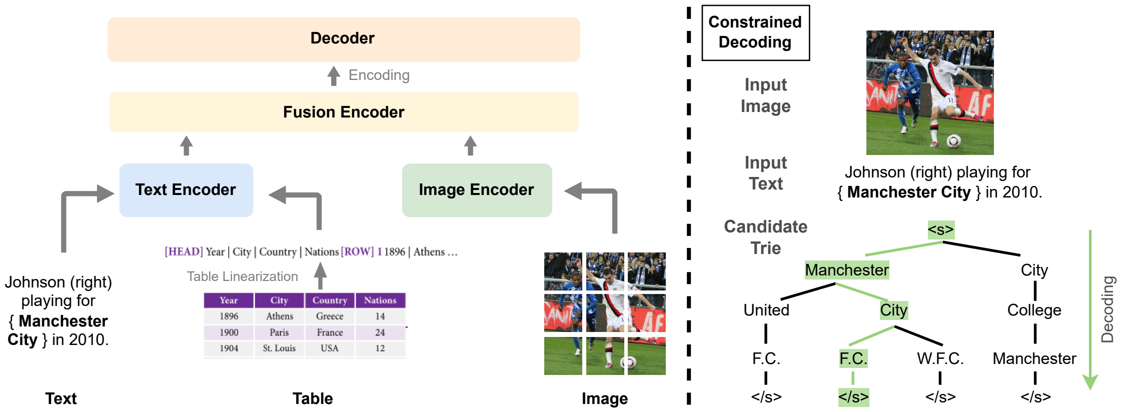

Decoder
Encoding
Fusion Encoder
Text Encoder
Image Encoder
Johnson (right) playing for { Manchester City } in 2010.
[HEAD] Year | City | Country | Nations [ROW] 1 1896 | Athens …
Table Linearization
Year | City | Country | Nations
1896 | Athens | Greece | 14
1900 | Paris | France | 24
1904 | St. Louis | USA | 12
Text
Table
Image
Constrained Decoding
Input Image
Input Text
Johnson (right) playing for { Manchester City } in 2010.
Candidate Trie
<s>
Manchester
City
United
City
College
F.C.
F.C.
W.F.C.
Manchester
</s>
</s>
</s>
</s>
Decoding

Figure 2: Overall GDMM architecture (left) and constrained decoding illustration (right).

Dataset	Modality	# Sentences / Tables Train	Valid	Test
GERBIL	L → L	18,448	4,791	19,614
WikiDiverse	VL → L	6,312	755	757
MELBench	VL → L	13,216	1,888	3,776
Squall	LU → L	9,028	2,246	-
SLSQL	LU → L	7,000	1,034	-

Table 2: Data statistics. Modalities include (L)anguage, (V)ision, and Tabl(U)r modality.

schema linking on Squall and SLSQL. All datasets are in English, and we summarize the data statistics in Table 1. We will release this benchmark.

We assume that the mention span is given across all datasets, thus (1) on GERBIL benchmark, we investigate textual entity disambiguation using the same candidate sets as in Le and Titov (2018) and De Cao et al. (2021). (2) On WikiDiverse, we investigate entity disambiguation performance with retrieved Top-10 candidates by Wang et al. (2022d) (3) On MELBench-Wikidata, we follow the original setting in which the given entity mention will be linked to its referent in the knowledge base. (4) On Squall, the given entity mention in the natural question will be linked to column names in the target table. (5) On SLSQL, the entity mention in the natural question will be linked to the column name within multiple tables.

Statistics for each individual EL task are shown in Table 2. Note that for the GERBIL benchmark, the training split refers to AIDA-train, the validation split refers to AIDA-dev, and the test split includes all test split as in (De Cao et al., 2021).

4 GDMM Model

We build a generative entity linking model that enables diverse-modal vision, language and table understanding and inference. We show how this can be achieved with a generative encoder-decoder structure.

4.1 Input Processor

As shown in Figure 2, our model can process inputs of three modalities, including texts, images, and tables. Formally, given a multimodal input $\{\mathbf{x}_i, \mathbf{v}_i, \mathbf{u}_i\}$, the input processor serves to encode the data and group the modalities as follows. (1) **Text** Given an input text $\mathbf{x}_i$, we first tokenize and embed it into a list of word vectors $\boldsymbol{x}_i$ following Devlin et al. (2019). (2) **Image** Given an input image $\mathbf{v}_i$, we first resize it to a fixed size and split it into patches, following Kim et al. (2021). (3) **Table** We follow the table representation proposed in TAPEX (Liu et al., 2022). The table is flattened and represented as $\mathbf{u}_i^* = \texttt{[head]}, col_1, \cdots, col_M, \texttt{[ROW]}, 1, cell_{11}, \cdots, cell_{1M}, \texttt{[ROW]} \cdots$, where [head] and [ROW] are special tokens denoting the beginning of table headers and rows, and the number after [ROW] is used to denote the row index. $cell_{ij}$ represents a cell in ith row and jth column. Then the table representation $\mathbf{u}_i^*$ will be tokenized and embedded into $\boldsymbol{u}_i$. Finally, given the multimodal input $\{\mathbf{x}_i, \mathbf{v}_i, \mathbf{u}_i\}$, our input processor outputs $\{\boldsymbol{x}_i \bigoplus \boldsymbol{u}_i, \boldsymbol{v}_i\}$, where $\bigoplus$ is a concatenation operator.

4.2 GDMM Model Architecture

Multimodal Encoder The multimodal encoder consists of an image encoder, a text encoder, and a fusion encoder, following previous work (Singh et al., 2022; Yang et al., 2022). The text encoder and image encoder use the same ViT architecture (Dosovitskiy et al., 2021) with different parameters. The text input and visual input $\{\boldsymbol{x}_i \bigoplus \boldsymbol{u}_i, \boldsymbol{v}_i\}$ are passed into the text encoder and vision encoder individually. The text hidden state vectors $\{\boldsymbol{h}_i^{LU}\}$ and the image embeddings $\{\boldsymbol{h}_i^{V}\}$ are then projected and concatenated into a single list. A fusion encoder is applied to the concatenated list, which allows cross-attention between the projected unimodal representations and fuses the two. The output is a list of hidden states $\{\boldsymbol{h}_i^{M}\}$. The multimodal encoder parameters are initialized with pre-trained FLAVA (Singh et al., 2022) parameters.

Decoder We exploit the transformer architecture (Vaswani et al., 2017) for the decoder. Previous study (Rothe et al., 2020) points out that combining models with same vocabulary has stronger overall performance, thus we initialize the decoder with the BERT (Devlin et al., 2019) pre-trained parameters.

Training and Inference The GDMM is trained with standard autoregressive objective, i.e., maximizing the output sequence likelihood $p_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{v}_i, \boldsymbol{u}_i)$ with respect to the model's parameters θ. We rank each candidate $\mathbf{c}_i^k \in \mathcal{C}_i$ by computing a score with an autoregressive formulation: $\text{score}(\mathbf{c}_i^k|\boldsymbol{x}_i, \boldsymbol{v}_i, \boldsymbol{u}_i) = p_\theta(\boldsymbol{y}_i^k|\boldsymbol{x}_i, \boldsymbol{v}_i, \boldsymbol{u}_i) = \prod_{j=1}^{N} p_\theta(y_j^k|y_{<j}^k, \boldsymbol{x}_i, \boldsymbol{v}_i, \boldsymbol{u}_i)$, where N is the number of tokens of $\mathbf{c}_i^k$. If the score is lower than a threshold θ, the prediction becomes "nil". The threshold will be decided by the development set.

Constrained decoding When the candidate set $\mathcal{C}_i$ is very large (e.g., the entire entity space $\mathcal{E}$), naturally, it is intractable to compute a score for every element. Thus we exploit Constrained Beam Search (Sutskever et al., 2014; De Cao et al., 2021), a tractable decoding strategy to efficiently search the valid entity space. It is tractable as the average time cost depends on beam size and the average length of entity representations (e.g. 6 BPE tokens on average for entities in Wikipedia KB), instead of the size of $\mathcal{C}_i$. An entity trie $\mathcal{T}_i$ for $\mathcal{C}_i$ will be created so that the output is limited to the target space. The constraint is defined as, for each node $t \in \mathcal{T}$, its children indicate all allowed continuations from the prefix traversing from root to t. For example, as shown in Figure 2, given four candidates `Manchester United F.C.`, `Manchester City F.C.`, `Manchester City W.F.C`, and `City College Manchester`, a candidate trie will be created as shown in the figure. The decoding will strictly follow the top-down order in the trie with a certain beam size.

4.3 Pre-training GDMM

Pre-training is critical to our architecture though the encoder and decoder are initialized with pre-trained weights because the mapping between the encoder and decoder are randomly initialized and they have not been pre-trained simultaneously with the encoders and the decoder.

Pre-training data A pre-training corpus is constructed from BLINK (Wu et al., 2020) and images in Wikipedia KB. BLINK is a commonly used corpus for textual entity linking pre-training, including 9M unique annotations of document-mention-entity triples from Wikipedia. Meanwhile, the images in Wikipedia KB are naturally linked to their respective entity names. The two together are well-suited for pre-training DMEL models. Aside from **text-only** BLINK, we construct **LV-paired** pre-training data by linking BLINK and Wiki-images. An image pool (Wiki-images) is collected from Wikipedia KB if the entity can be linked to mentions in BLINK. The image pool contains 797,436 downloaded images of 495,149 entities in Wikipedia KB. We then randomly attach an image of the target entity, if exists, to each mention in BLINK. In total, the LV-paired pre-training data includes 5,445,264 mentions and 678,385 distinct images in the training set, and 5,816 mentions and 5,414 images in the development set.

Pre-training details We pre-train GDMM on text-only BLINK and LV-paired pre-training data in two stages. Note that not all the BLINK entities appear in Wiki-images. There are over 2.5M BLINK mentions not covered by LV-paired pre-training data. To fully leverage the BLINK annotations, we first pre-train on text-only BLINK and then pre-train on the LV-paired data. With text-only BLINK, we freeze parameters in the image encoder and fusion layers and only update parameters in the text encoder and decoder. In the second stage, all the parameters are updated.

4.4 Unified Learning

Upon the pre-trained model, one straightforward strategy for downstream tasks is single-task finetuning. We take one step further and investigate unified learning. Specifically, we'll investigate (1) single-task finetuning (ST-F), which refers to

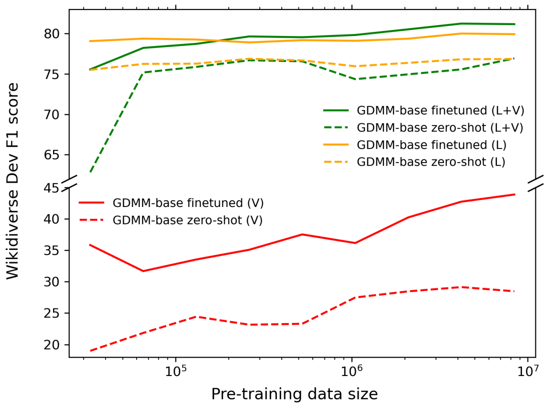


Figure 3: The effect of the number of paired BLINK and Wiki-images pre-training data on GDMM-base.

finetuning on individual tasks; (2) multi-task finetuning (MT-F) which combines the mixed training data of all datasets (Raffel et al., 2022); and (3) multitask fine-tuning with prefixes (MT-FP), where we prepend task-specific prefixes like “entity linking” and “schema linking” to the input context.

5 Experiments

Model Variants We primarily report results on two model variants: **GDMM-base** where the decoder is initialized with BERT-base parameters, and **GDMM-large** where the decoder is initialized with BERT-large parameters. To investigate which modality provides dominant information for visual-text entity linking, three configurations are explored: **L+V** where both visual and textual information are given, **L** where only textual input are given, and **V** where only image are given. We report experimental results with a **single generic model** for the three modality settings. It is achieved by randomly masking out one modality during training. Implementation details are in Appendix B.

5.1 Results

Pre-training The pre-training of GDMM consists of two stages, text-only pre-training and LV-paired pre-training. The pre-training performance is investigated with two methods: zero-shot or fine-tuned on WikiDiverse. Zero-shot refers to directly evaluating on WikiDiverse without training, while fine-tuned refers to further fine-tuning on the WikiDiverse. The first-stage pre-trained checkpoint is directly evaluated on WikiDiverse with only text information (note that WikiDiverse is a dataset with both image and text input), achieving 75.43 zero-shot F1 score. It greatly outperform the baseline model in (Wang et al., 2022d) by 4.36 F1 score, even though only text information are leveraged. The evaluation demonstrates that pretraining on BLINK builds a strong foundation for the proposed model. After that, we investigate the effect of paired pre-training data size in the second stage and visualize it in Figure 3. It shows that the pre-training data size has a positive effect on inference with both image and text modality (L+V) and with only image modality (V). Text-only (L) performance is not affected even though the visual modality is introduced in the second pre-training stage.

Experimental results on DMEL benchmark The experimental results of the proposed GDMM on DMEL are shown in Table 3. We compare visual-language entity disambiguation result on WikiDiverse with LXMERT (Wang et al., 2022d), and visual-language entity linking performance on MELBench with Gan et al. (2021a). GDMM achieves better performance on both datasets, especially on MELBench. GDMM strikingly improves the F1 score by over 31%, demonstrating the effectiveness of the proposed architecture.

The schema linking performance is evaluated on Squall and SLSQL. For schema linking, we compare our model with the baseline model GENRE (De Cao et al., 2021), as it has competitive performance in entity disambiguation. For a fair comparison, we fine-tune GENRE with their pretrained checkpoint on BLINK and investigate two options, one without a flattened table (GENRE) and another with the flattened table (GENRE+) where the table content is leveraged identically as in GDMM. Only experimental results for GENRE+ are reported in Table 3 since it has better performance than GENRE. The fact that the flattened table has better performance demonstrates the effectiveness of the table representation. Detailed results can be found in Appendix C.

Unified learning As mentioned in Section 4.4, we report unified learning results for ST-F, MT-F, and MT-FP in Table 4. To confirm whether the pre-trained checkpoints build a competitive foundation for visual-language entity linking, zero-shot (ZS) performance is also reported in the same table. The pre-trained checkpoint is competitive because the zero-shot performance (i.e. ZS column) outperforms previous state-of-the-art fine-tuned re-

Data	Task	Modality	Previous SOTA	GDMM-base	GDMM-large
GERBIL	ED	L → L	**88.8** (De Cao et al., 2021)	86.11±0.24	82.57±0.22
WikiDiverse	VED	LV → L	71.07 (Wang et al., 2022d)	**79.10**±0.35	78.69±0.33
MELBench	VED	LV → L	40.5 (Gan et al., 2021a)	68.01±0.75	**72.41**±0.65
Squall	SL	LU → L	82.10±2.41 (GENRE+)	**89.69**±0.77	89.12±1.03
SLSQL	SL	LU → L	82.80 (GENRE+)	81.48±1.06	**84.43**±0.92
Avg.			72.93	80.88	81.44

Table 3: Benchmark results on DMEL data

Dataset	**GDMM-base**			
	ZS	ST-F	MT-F	MT-FP
GERBIL	84.00	**93.75** ±0.26	93.63 ±0.14	93.56±0.52
WikiDiverse	76.92	**80.97**±0.39	80.02±0.29	80.65±0.35
MELBench	54.76	**67.41**±0.97	63.64±2.04	65.64±1.44
SQUALL	47.52	**89.69**±0.77	88.00±1.31	88.37±0.99
SLSQL	30.92	81.48±1.06	83.59±1.90	**83.60**±0.85
Avg.	58.82	**82.66**	81.78	82.36

Table 4: Unified learning results. Best scores and second-best scores are highlighted in **Bold** and underlined. Variance does not apply to zero-shot F1 scores because the pre-trained checkpoint is unique. We report results on the development set.

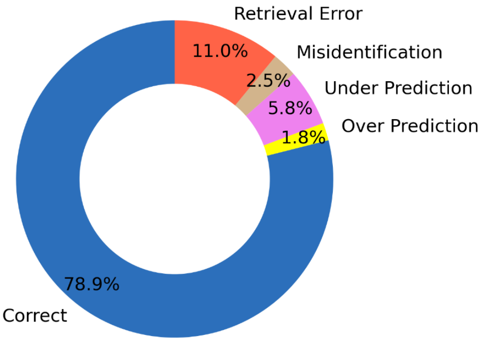


Figure 4: Error breakdown on WikDiverse validation set.

sults for WikiDiverse in Table 8 and MELBench in Table 9. On average, ST-F and MT-FP achieve the best and the second-best performance, with a small gap between the two. It is expected that ST-F achieves the best performance as each fine-tuned model is able to fit the target dataset distribution. Considering ST-F trains five models while MT-FP trains a single model, the competitive MT-FP performance suggests that model efficiency can be achieved at the cost of a minor performance drop, that is, 0.30 average F1 score drop for GDMM-base. Additionally, the fact that MT-FP constantly outperforms MT-F aligns with previous findings that task-specific prefixes is effective in informing the model of the target tasks (Dong et al., 2019; Raffel et al., 2022).

5.2 Error Analysis

Figure 4 shows the error breakdown on WikiDiverse. The errors are divided into four categories: retrieval error where the target entity is not in the candidates; misidentification where the prediction does not match the ground truth entity; under predict where the model predicts "nil" and the ground truth entity is not "nil"; over prediction where the ground truth entity is "nil".

Representative error examples are presented in Table 5 for (a) retrieval error, (b) misidentification, (c) over prediction, and (d) under prediction. Error type (a) contributes to over half of the errors, emphasizing the need for a good retriever. It cannot be addressed by our model, because the ground truth entity is not in the set. Example (b) is due to candidate confusion, as Cape Canaveral Air Force Station is a previously used name for Cape Canaveral Space Force Station from 1974 to 1994 and from 2000 to 2020. Such errors indicate the necessity for a coreference system at inference time. The over-prediction example as shown in (c) calls for a better discrimination strategy for plausible candidates. Example (d) is a challenging example, which asks future models to possess more profound prior knowledge.

Table 6 further shows four types of errors for schema linking, name ambiguity, inference difficulty, prime key confusion, and unknown strings. (a) Name ambiguity is a common challenge for schema linking especially when the column names have overlapped tokens. (b) Sometimes the model fails to make inferences on subtle entity expressions. (c) Another common error type for schema linking is the confusion in prime keys, as the prime key "pet id" is shared by multiple tables in the example. (d) Another challenge is unknown strings or composite tokens since it is usually intractable to recover the original expression from those mentions.

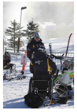
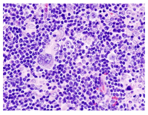
ID	(a)	(b)	(c)	(d)
Image				
Text	Australians before the **competition** started for the day at the bottom of the **hill**	spacecraft aboard lifts off at **Cape Canaveral Air Force Station** at 6:01 p.m	**pilgrims** at Mecca's Grand Mosque in 2008	**Histopathologic image** of Hodgkin's lymphoma.
GT	Nor-Am Cup ; Copper Mountain (Colorado)	Cape Canaveral Space Force Station	nil	Histopathology
Pred (L+V)	Competition ; The Hill (Film)	Cape Canaveral Air Force Station	Pilgrims (Plymouth Colony)	nil

Table 5: Examples for (a) retrieval error, (b) misidentification, (c) over prediction, and (d) under prediction.

ID	Error Type	Text	GT	Prediction
(a)	Column name ambiguity	show the **name** and the release year of the song by the youngest singer .	singer # song name	singer # name
(b)	Lack of inference	which model be **lighter** than 3500 but not build by the ' Ford Motor Company ' ?	cars data # weight	model list # model
(c)	Prime key confusion	find the **id** of the pet own by student whose last name be ' Smith ' .	has pet # pet id	pets # pet id
(d)	Unknown strings	... list the car **makeid** and make name .	car names # make id	car names # make

Table 6: Case study on schema linking errors. The hash symbol connects the table name and column name.

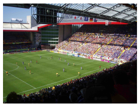
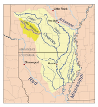
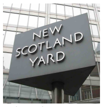
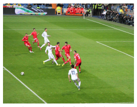
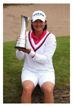
ID	(a)	(b)	(c)	(d)	(e)
Image					
Text	**Australia** and **Japan** at the 2006 World Cup	... around the **Little Missouri** (highlighted) and Caddo rivers.	... iconic sign at current Metropolitan Police **headquarters**.	Real Madrid players including **Ronaldo** go for the ball following a foul.	**Tseng** with the trophy
Pred(L+V) =GT	Australia national soccer team ; Japan national football team	Little Missouri River (Arkansas)	Scotland Yard	Cristiano Ronaldo	Yani Tseng
Pred (L)	Australia ; Japan	Little Missouri River Bridge	Headquarters	Ronaldo (Brazilian Footballer)	Sam Tseng

Table 7: Ablation on when the visual content is indispensable for entity disambiguation.

5.3 Ablation Study

We further discuss the indispensability of information within each modality through an ablation study on WikiDiverse. Three evaluations are conducted under settings: (1) L+V where both text and image are leveraged; (2) L where only texts are used for prediction; (3) V where only images are used for prediction. The experimental results are reported in Table 8. It shows each model achieves the best performance with both text and image modalities (L+V). While textual content provides the most inference clues, visual content provides complementary information. Additionally, the proposed model GDMM outperforms LXMERT with a single generic model trained for various configurations, while three modality-specific models for each configuration (L+V, L, V) are trained in LXMERT. This observation demonstrates the effectiveness of the proposed model for various modality configurations.

Table 7 shows several misinformation examples where the model fails without visual information. In (a), the image of a soccer stadium provides extra semantics when the model misses the semantic indicator from "the 2006 World Cup." The image in example (b) is in Wiki-KB and has been used for pre-training. Its visibility during the pre-training makes the image a strong indicator of the target entity. The optical character "NEW SCOTLAND YARD" in example (c) is indispensable for the mention to be correctly identified. Without the

Method	F1		
	L+V	L	V
LXMERT (Wang et al., 2022d)	71.07	63.65	40.16
GDMM-base	**79.10**±0.35	76.79±0.32	**40.59**±2.32
GDMM-large	78.69±0.33	**77.05**±0.29	37.65±0.25

Table 8: Ablation study on WikiDiverse.

optical features, inferring the mention of “headquarters” is challenging given the ambiguous text. Examples (d) and (e) emphasize the dependence on facial features for entity disambiguation. In the absence of the image in (d), it is impossible to disambiguate between “Cristiano Ronaldo” and “Ronaldo (Brazilian Footballer)” as both players served Real Madrid.

6 Related Works

Textual Entity Linking Early entity linking researches (Hoffart et al., 2011; Daiber et al., 2013) reply on probabilistic approaches, based on textual similarity and corpus occurrence. A more recent line of research is neural networks based retrieval-reranking approaches (El Vaigh et al., 2020; Zhang et al., 2022; Mrini et al., 2022), which first retrieve top candidates given the input text, and then score each candidate with semantic similarity or correlation. End-to-end entity linking models (Broscheit, 2019; Martins et al., 2019; El Vaigh et al., 2020) approach this problem by directly detecting the entity mentions and linking them to their corresponding entities in the KB. For example, autoregressive entity linking models (De Cao et al., 2021; De Cao et al., 2021; Petroni et al., 2021; Mrini et al., 2022) formulate entity linking as a language generation problem using an encoder-decoder model.

Textual-Visual Entity Linking The growing trend towards multimodality significantly advanced research in multimodal entity linking. Due to the difficulty in collecting and cleaning multimodal entity linking data, previous researchers limit their attention to a specific domain such as social media data (Adjali et al., 2020b,a; Moon et al., 2018; Gan et al., 2021a) and news domain(Zheng et al., 2022; Wang et al., 2022d), or a limited scope like person and organization recognition (Gan et al., 2021a; Cui et al., 2021). Previous work (Wang et al., 2022d) represents each entity with one image, which limits the visual expression of entities. We overcome this limitation by pre-training GDMM with multiple images per entity to obtain diverse visual representations.

Tabular Schema Linking Schema linking (Guo et al., 2019; Wang et al., 2020) is an instance of entity linking in the context of linking to the relational database schema. Previous research shows that good schema linking (Liu et al., 2021; Katsakioris et al., 2022; Shi et al., 2020; Lei et al., 2020; Chen et al., 2020) can substantially improve downstream tasks such as Text-to-SQL parsing. However, entity mentions in existing benchmarks such as Spider (Yu et al., 2018) can almost exactly match the corresponding schema entities (Chen et al., 2020). Therefore, current Text-to-SQL semantic parsers normally address this problem with string-matching heuristics (Chen et al., 2020) or embedding matching modules (Chen et al., 2020; Wang et al., 2022a; Guo et al., 2019; Wang et al., 2020). However, due to the diversity and ambiguity in natural language mentions, such heuristics are hard to generalize to new domains (Chen et al., 2020; Wang et al., 2022a).

Multimodal Models Multimodal models have attracted increasing attention in computer vision and natural language processing communities. Recent transformer-based approaches (Kim et al., 2021; Radford et al., 2021; Singh et al., 2022) that leverage the attention between the visual and textual embeddings manifest the effectiveness of the attention mechanism. However, the proposed learning objectives are usually limited to predefined scopes, such as text-image matching or alignment (Xu et al., 2021; Radford et al., 2021; Biten et al., 2022; Ho et al., 2022; Li et al., 2022; Yang et al., 2022; Huang et al., 2023), semantic segmentation, object detection, classification (Xu et al., 2022; Guo et al., 2022; Assran et al., 2022), and masked language modeling (Li et al., 2020; Ni et al., 2022; Tong et al., 2022; Appalaraju et al., 2021). Instead, we proposed a generic generative model that is open to diverse downstream tasks. Additionally, GDMM differs from previous generative multimodal models(Li et al., 2023; Wang et al., 2021) in that GDMM can process and comprehend information from heterogeneous instead of a single source; GDMM differs from VL-T5 (Cho et al., 2021) and others (Wang et al., 2022b; Bao et al., 2022; Wang et al., 2022c) in that GDMM enables thorough encoding for each modality, instead of discretizing visual content.

7 Conclusion

In this paper, a novel DMEL problem is formulated, which links the entity mention within heterogeneous information to a defined KB. A generic DMEL dataset is built covering diverse EL tasks. We propose a unified generative model for DMEL, GDMM. Comprehensive experiments are conducted over the DMEL dataset. Experimental results show that the proposed GDMM outperforms state-of-the-art models on almost each individual EL task.

Broader Impact In contrast to previous work (Pan et al., 2022; OpenAI, 2022) that only leverage textual content, the proposed model has the potential to deal with misinformation (text only as a data source might be prone to misinformation or fake context/information). This research will lead to a clearer understanding of misinformation issues and encourage better leverage of multimodal information.

Limitations

GDMM establishes a compelling starting point for DMEL research. In spite of this, the proposed approach has several shortcomings. First, GDMM currently generates entity name within the entity candidate set, however, we saw how retrieval errors limit entity linking performance. Thus, how to work collectively with the retrieval system to diminish errors takes appropriate action. Second, how to handle large tables still remains under-explored. It is infeasible to represent a huge database with the table flattening technique. In practice, it is possible to filter out less likely candidates to compress the search space, but a more promising approach is to represent the table more efficiently.

GDMM also enables studies on more diverse-modal tasks. New tasks can be easily framed based on the proposed architecture, such as visual question answering, grounded generation, and diverse-modal commonsense reasoning. We believe that with more follow-up work on diverse tasks, this approach will turn out to be a more comprehensive generative diverse-modal framework.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020a. Multimodal entity linking for tweets. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I*, pages 463–478, Berlin, Heidelberg. Springer-Verlag.

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020b. Building a multimodal entity linking dataset from tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4285–4292, Marseille, France. European Language Resources Association.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 973–983.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. 2022. Masked siamese networks for label-efficient learning.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. 2022. Latr: Layout-aware transformer for scene-text vqa. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16527–16537.

Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 423–433, Sofia, Bulgaria. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Sanxing Chen, Aidan San, Xiaodong Liu, and Yangfeng Ji. 2020. A tale of two linkings: Dynamically gating between schema linking and structural linking for text-to-SQL parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2900–2912, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*.

Claire Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. 2021. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1374–1384.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124, New York, NY, USA. Association for Computing Machinery.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Highly parallel autoregressive entity linking with discriminative correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7662–7669, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Cheikh Brahim El Vaigh, François Torregrossa, Robin Allesiardo, Guillaume Gravier, and Pascale Sébillot. 2020. A correlation-based entity embedding approach for robust entity linking. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 949–954.

D. A. Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1:1–1:15.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-sql evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360. Association for Computational Linguistics.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021a. Multimodal entity linking: A new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 993–1001, New York, NY, USA. Association for Computing Machinery.

Yujian Gan, Xinyun Chen, Qiuping Huang, Matthew Purver, John R. Woodward, Jinxia Xie, and Pengsheng Huang. 2021b. Towards robustness of text-to-SQL models against synonym substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2505–2515, Online. Association for Computational Linguistics.

Yujian Gan, Xinyun Chen, and Matthew Purver. 2021c. Exploring underexplored limitations of cross-domain text-to-SQL generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8926–8931, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and D. Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. In *ACL*.

Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. 2022. Visual attention network. *arXiv preprint arXiv:2202.09741*.

Chih-Hui Ho, Srikar Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. 2022. Yoro-lightweight end to end visual grounding. In *ECCV 2022 Workshop on International Challenge on Compositional and Multimodal Perception*.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. 2023. Nlip: Noise-robust language-image pretraining. In *AAAI 2023*.

Miltiadis Marios Katsakioris, Yiwei Zhou, and Daniele Masato. 2022. Entity linking in tabular data needs the right attention.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

Wenqiang Lei, Weixin Wang, Zhixin Ma, Tian Gan, Wei Lu, Min-Yen Kan, and Tat-Seng Chua. 2020. Re-examining the role of schema linking in text-to-SQL. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6943–6954, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *AAAI 2023*.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. TAPEX: Table pre-training via learning a neural SQL executor. In *International Conference on Learning Representations*.

Qian Liu, Dejian Yang, Jiahui Zhang, Jiaqi Guo, Bin Zhou, and Jian-Guang Lou. 2021. Awakening latent grounding from pretrained language models for semantic parsing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1174–1189, Online. Association for Computational Linguistics.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2019. Joint learning of named entity recognition and entity linking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 190–196, Florence, Italy. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Khalil Mrini, Shaoliang Nie, Jiatao Gu, Sinong Wang, Maziar Sanjabi, and Hamed Firooz. 2022. Detection, disambiguation, re-ranking: Autoregressive entity linking as a multi-task problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1972–1983, Dublin, Ireland. Association for Computational Linguistics.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 1–18, Berlin, Heidelberg. Springer-Verlag.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt/`.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2022. ContraQA: Question answering under contradicting contexts.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

P. J. Price. 1990. Evaluation of spoken language systems: the ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990.*

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Tianze Shi, Chen Zhao, Jordan Boyd-Graber, Hal Daumé III, and Lillian Lee. 2020. On the potential of lexico-logical alignments for semantic parsing to SQL queries. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1849–1864, Online. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Yasufumi Taniguchi, Hiroki Nakayama, Takahiro Kubo, and Jun Suzuki. 2021. An investigation between schema linking and text-to-sql performance. *ArXiv*, abs/2102.01847.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ruben Verborgh, Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. Gerbil – benchmarking named entity recognition and linking consistently. *Semant. Web*, 9(5):605–625.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7567–7578, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. 2021. Large-scale self- and semi-supervised learning for speech translation.

Lihan Wang, Bowen Qin, Binyuan Hui, Bowen Li, Min Yang, Bailin Wang, Binhua Li, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022a. Proton: Probing schema linking information from pre-trained language models for text-to-sql parsing. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pages 1889–1898, New York, NY, USA. Association for Computing Machinery.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022c. Image as a foreign language: Beit pretraining for all vision and vision-language tasks.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022d. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.

Ledell Yu Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *EMNLP*.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha

Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *CVPR 2022*.

Yi Yang, Ozan Irsoy, and Kazi Shefaet Rahman. 2018. Collective entity disambiguation with structured gradient tree boosting. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 777–786, New Orleans, Louisiana. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2022. EntQA: Entity linking as question answering. In *International Conference on Learning Representations*.

Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. 2022. Visual Entity Linking via Multi-modal Learning. *Data Intelligence*, 4(1):1–19.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

A DMEL Selection

The DMEL benchmark's datasets were carefully chosen after conducting extensive research on publicly available datasets. The DMEL benchmark includes five datasets: GERBIL, WikiDiverse, MELBench, Squall, and SLSQL. Each of these datasets is necessary and represents the best option for a comprehensive evaluation for Diverse-Modal Entity Linking. The selected datasets best align with the DMEL problem, among more than 20 datasets we looked into. The reasons why other datasets are excluded in the benchmark are: i) most dataset collected from social media cannot be reproduced because some of the data are no longer accessible. This category includes Twitter-MEL (Adjali et al., 2020b), SnapCaptionsKB(Moon et al., 2018); ii) Dataset is not publicly available (Zheng et al., 2022); iii) Annotated dataset for schema linking is limited. Existing work that investigates entities in tables are Text-to-SQL datasets. However, annotations for schema linking are not available, such as Spider (Yu et al., 2018), Spider-Syn (Gan et al., 2021b), Spider-DK (Gan et al., 2021c), WikiSQL(Zhong et al., 2017), ATIS (Price, 1990), Freebase917 (Cai and Yates, 2013), and WikiTableQuestions (Pasupat and Liang, 2015). Furthermore, there is no overlap between the datasets. For tabular datasets, Squall is built upon WikiTableQuestions, while SLSQL is based on the Spider text-to-SQL dataset. Lastly, GENRE is widely recognized as the standard dataset for the task of textual entity linking.

B Implementation Details

For every dataset in DMEL dataset, the fine-tuning procedure runs for 5 epochs with a batch size of 16. For both pre-training and fine-tuning, the learning rate is 3×10^{-5}, with a linear scheduler with 0.1 warmup ratio. Fine-tuning takes 5 hours for WikiDiverse, 2 for MELBench-Wiki, 1 for Squall, and 2 for SLSQL.

We run each setting 5 times and report the mean and variance unless stated otherwise (except the zero-shot setting when evaluated with the pre-trained checkpoint, since there is no randomness with the pre-trained checkpoint). One MELBench, since the dataset split is not given along with the released MELBench-Wikidata data, we randomly split the dataset according to their split statistics and repeat the experiment 5 times to get average evaluation metrics. On Squall, we reported a 5-fold cross-validation result following the released split. For Squall and SLSQL, all the hyperparameters are tuned on the training set since there is no test set.

Additionally, Wikipedia images are collected through hyperlinks shared by Wang et al. (2022d) at `https://github.com/wangxw5/wikidiverse`. Annotations on SLSQL are adapted from Spider, excluding train_others.json that are from Restaurants, GeoQuery, Scholar, Academic, IMDB, and Yelp prepared by (Finegan-Dollak et al., 2018). For GERBIL benchmark results, we report average F1 scores on six test sets, including Aida-test, MSNBC-test AQUAINT-test, ACE2004-test, WNED-CWEB-test, and WNED-WIKI-test following De Cao et al. (2021).

C Detailed Experimental Results

Experimental results for each individual dataset are shown in this section. Specifically, Table 9 shows results for MELBench, Table 10 shows the experimental result for Squall, and Table 11 shows the experimental result for SLSQL.

D Domain Adaption Results

We investigate zero-shot performance in unseen domains on the WikiDiverse dataset. Specifically, we choose the five domains as seen domains, including politic, crime, sports, entertainment, and technology, and the rest five domains as the unseen domains, including disaster, health, economy, weather, and education. Training data includes instances from the seen domain, and the instances from the unseen domains are randomly split into validation and test set. Note that the data used for this experiment is from the WikiDiverse training and validation set, the data in the test set are excluded.

The domain adaption result on WikiDiverse is shown in Table 12. These experiments facilitate studying knowledge transfer between seen and unseen domains. The experiment results show that (a) pre-training is indispensable for new domains as it provides profound prior knowledge for the MEL task in general; (b) knowledge learned from seen domains can indeed transfer to the unseen domain as the average F1 score improves by 3.61 percentage points.

Method	F1	
	Top-1	Top-10
MELBench (Gan et al., 2021a)	40.5	69.6
GDMM-base	68.01±0.75	73.31±0.77
GDMM-large	**72.41**±0.65	**76.34**±0.78

Table 9: Results of entity linking on MELBench

Method		F1
GENRE		75.92±4.29
GENRE+		82.10±2.41
GDMM-base	Zero-shot	47.52±1.06
	Finetuned	**89.69**±0.77
GDMM-large	Zero-shot	49.14±1.26
	Finetuned	89.12±1.03

Table 10: Results of schema linking on Squall. GENRE+ denotes augment table representations to the input text as described in Section 4.1

Method		F1
GENRE		70.41
GENRE+		82.80
GDMM-base	Zero-shot	30.92
	Finetuned	81.48±1.06
GDMM-large	Zero-shot	28.44
	Finetuned	**84.43**±0.92

Table 11: Results of schema linking on SLSQL

Domain	F1		
	FT w/o PT	PT	FT with PT
Health	40.37	76.19	82.48
Weather	39.80	79.79	78.84
Economy	46.67	78.95	84.17
Disaster	34.67	78.81	81.04
Education	40.23	73.49	81.48
Overall	39.17	78.05	81.66

Table 12: Evaluation on unseen domains in WikiDiverse with GDMM-base. FT and PT stand for fine-tuning on seen domains and pre-training.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☑ A2. Did you discuss any potential risks of your work?
Section 5 and Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 3, 4, 5

☑ B1. Did you cite the creators of artifacts you used?
Sections 3, 4, 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix B

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix B

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 and Appendix B

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3 and Appendix B

C ☑ Did you run computational experiments?

Section 5 and Appendix B

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5 and Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 and Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5 and Appendix B

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.