MESED: A Multi-Modal Entity Set Expansion Dataset with Fine-Grained Semantic Classes and Hard Negative Entities

Yangning Li[1,2*], **Tingwei Lu**[1*], **Hai-Tao Zheng**[1,2†], **Yinghui Li**[1†],
Shulin Huang[1], **Tianyu Yu**[1], **Jun Yuan**[3], **Rui Zhang**[3]
[1]Shenzhen International Graduate School, Tsinghua University
[2]PengCheng Laboratory
[3]Huawei Noah's Ark Lab
{yn-li23,ltw23}@mails.tsinghua.edu.cn

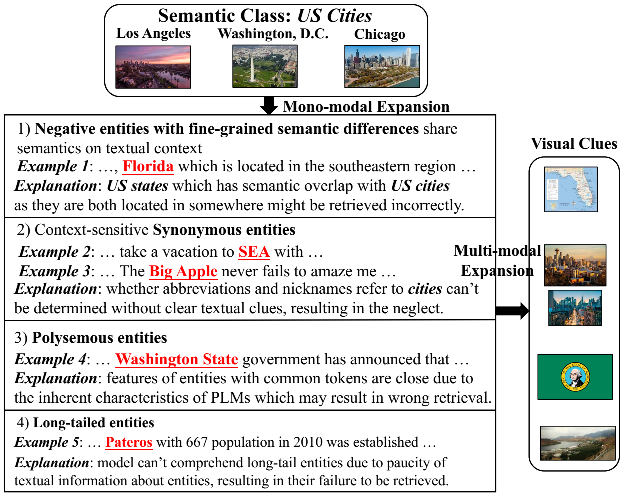
Abstract

The Entity Set Expansion (ESE) task aims to expand a handful of seed entities with new entities belonging to the same semantic class. Conventional ESE methods are based on mono-modality (i.e., literal modality), which struggle to deal with complex entities in the real world such as (1) Negative entities with fine-grained semantic differences. (2) Synonymous entities. (3) Polysemous entities. (4) Long-tailed entities. These challenges prompt us to propose novel Multi-modal Entity Set Expansion (MESE), where models integrate information from multiple modalities to represent entities. Intuitively, the benefits of multi-modal information for ESE are threefold: (1) Different modalities can provide complementary information. (2) Multi-modal information provides a unified signal via common visual properties for the same semantic class or entity. (3) Multi-modal information offers robust alignment signals for synonymous entities. To assess model performance in MESE, we constructed the MESED dataset which is the first multi-modal dataset for ESE with large-scale and elaborate manual calibration. A powerful multi-modal model MultiExpan is proposed which is pre-trained on four multimodal pre-training tasks. The extensive experiments and analyses on MESED demonstrate the high quality of the dataset and the effectiveness of our MultiExpan, as well as pointing the direction for future research. The benchmark and code are public at https://github.com/THUKElab/MESED.

Introduction

The Entity Set Expansion (ESE) task aims to expand a handful of seed entities with new entities belonging to the same semantic class based on the given candidate entity vocabulary and corpus(Zhang et al. 2020; Li et al. 2022a). For example, given {*Washington D.C.*, *Chicago*, *Los Angeles*}, ESE tries to retrieve other entities with the target semantic class `US Cities`, such as *New York*, *NYC*, *Boston*. ESE plays a significant role in knowledge mining and benefits a variety of downstream NLP and IR applications (Chen, Cafarella, and Jagadish 2016; Li et al. 2023b).

Conventional ESE methods are based on mono-modality (i.e., literal modality), which typically suffer from limited information and sparse representation. Taking expanding `US Cities` as an example, the mono-modal ESE methods struggle to deal with complex entities in the real world from the following perspectives:



1) **Negative entities with fine-grained semantic differences** share semantics on textual context
Example 1: …, **Florida** which is located in the southeastern region …
Explanation: ***US states*** which has semantic overlap with ***US cities*** as they are both located in somewhere might be retrieved incorrectly.

2) Context-sensitive **Synonymous entities**
Example 2: … take a vacation to **SEA** with …
Example 3: … The **Big Apple** never fails to amaze me …
Explanation: whether abbreviations and nicknames refer to ***cities*** can't be determined without clear textual clues, resulting in the neglect.

3) **Polysemous entities**
Example 4: … **Washington State** government has announced that …
Explanation: features of entities with common tokens are close due to the inherent characteristics of PLMs which may result in wrong retrieval.

4) **Long-tailed entities**
Example 5: … **Pateros** with 667 population in 2010 was established …
Explanation: model can't comprehend long-tail entities due to paucity of textual information about entities, resulting in their failure to be retrieved.

Figure 1: An example of tricky entities that a mono-modal ESE model cannot handle.

- **Negative entities with fine-grained semantic differences** refer to entities that belong to the same coarse-grained semantic class as target class. These entities share semantics on textual context and are consequently challenging to be differentiated in detail. For instance, when expanding `US Cities`, it's inevitable to expand entities with the same parent class (i.e., `US Location`), such as *Florida* and *Texas* that are also located in the US.
- **Synonymous entities** mean entities have a variety of aliases. The ESE model can readily understand common aliases, while failing to comprehend these context-sensitive aliases (Henriksson et al. 2014; Schumacher and Dredze 2019) such as abbreviations and nicknames, since ascertaining the meaning of them necessitates explicit textual cues. For example, *SEA* only means Seattle in certain contexts, potentially leading to the omission of its retrieval.
- **Polysemous entities**, which stand for possible ambigu-

[*]These authors contributed equally.
[†]Corresponding author

ity of a textual mention referring to multiple entities. Since pre-trained language models learn semantics through word co-occurrence (Kenton and Toutanova 2019; Lauscher et al. 2020), entities comprising the same tokens are inherently closer. For example, the L2 distance from *Washington, D.C.* to *Washington State* is instead smaller than the distance to many other cities like *Austin* (8.89 vs. 10.02 we measured). As a result, entities merely with the same textual tokens may be wrongly retrieved.

- **Long-tailed entities** represent low-frequency entities in the corpus, such as obscure place names. Due to the inadequate textual description, the representation of these entities is frequently too sparse, posing a challenge to their retrieval.

The aforementioned situations lead to the advent of **Multi-modal Entity Set Expansion (MESE)**, where we integrate information from multiple modalities to represent entities and expand them to target semantic classes.

MESE can overcome the limitations of mono-modal approaches by leveraging multiple sources of information. The benefits of MESE include the following: Firstly, multi-modal information can complement the information provided by texts (especially for short texts), thereby enhancing model to comprehensively understand entities. Secondly, multi-modal information can serve as a cohesive signal that unites semantic classes based on shared visual properties or characteristics. For instance, when dealing with `Comic Book Characters`, the background and style of images can serve as uniform features of the comic book characters, distinguishing them from hard negative semantic class `Movie Characters`. Third, multi-modal information can facilitate the resolution of polysemous entities and provide clues for the alignment of synonymous entities. In addition, we argue that multi-modal information is particularly beneficial to rarely used synonymous entities or long-tail entities, as entities of lower frequencies tend to be more concrete concepts with stable visual representations.

Regrettably, despite the availability of diverse multi-modal data types (Li et al. 2023a; Yu et al. 2023a,c; Cheng et al. 2023a,b,c), there is currently no multi-modal dataset structured based on fine-grained semantic classes. To address this gap, we have constructed a large-scale, manually annotated MESE dataset called MESED, comprising 14,489 entities sourced from Wikipedia and 434,675 image-sentence pairs. To the best of our knowledge, MESED is the first multi-modal dataset for ESE with large-scale and elaborate manual calibration. MESED features several elements to accentuate the challenges of ESE. Firstly, we meticulously crafted a semantic class schema that consists of 26 coarse-grained and 70 fine-grained classes, with fine-grained classes that are mutually ambiguous (e.g., *Chinese actors* versus *US actors*) being assigned as hard negative classes for each other. Furthermore, synonymous and polysemous entities are added to amplify confusion between entities. Additionally, to evaluate models' capability in comprehending sparse entities, uncommon semantic classes were deliberately included.

In experiments, conventional text-based models, as well as emerging GPT-3.5, and various visual and multi-modal baseline models are evaluated. We also propose a powerful multi-modal model MultiExpan trained with four self-supervised multi-modal pre-training tasks that we designed, including masked entity prediction, contrastive learning, clustering learning, and momentum distillation.

To summarize, the main contributions are as follows:

- We present a novel Multi-modal Entity Set Expansion (MESE) task, which expands entities in multiple modalities.
- We first release a large-scale human-annotated MESE dataset called MESED, which is challenging as its fine-grained semantic classes and ambiguous candidate entities.
- We provide strong multi-modal baseline models MultiExpan and explore diverse self-supervised pre-training objectives for representation learning of multi-modal entities.
- Extensive experiments demonstrate the effectiveness of our MultiExpan and provide direction for future research.

Task Formulation

Definition 1 ***Multi-modal Entity Set Expansion (MESE).*** The inputs of MESE are a small set $S = \{e_1, e_2, ..., e_k\}$ that contains several seed entities describing a certain semantic class and a vocabulary V of candidate entities. Besides, a corpus D containing the multi-modal contexts $\{e_i, (t_1^i, v_1^i), ..., (t_n^i, v_n^i)\}$ for each entity e_i is given, in which t_n^i is a sentence comprising e_i and (t_n^i, v_n^i) forms an image-sentence pair. It is of note that arbitrary modality may be lacking in a given context.

Dataset Construction

In this section, we demonstrate the MESED construction procedure. Several factors, including the coverage and ambiguity of semantic classes, as well as the relevance between images and entities are considered to ensure the quality of MESED.

Data Collection

There are two ways to construct a multi-modal ESE dataset. The first straightforward approach is to first collect the image-sentence pairs and label the entities in the sentences. Then, for each semantic class, human annotators traverse the entire large-scale entity vocabulary once to pick up the corresponding entities. Although plenty of public datasets are available with massive image-sentence pairs, the labour cost of such a bottom-up manner is prohibitive. We therefore adopt the more practical top-down approach to constructing MESED. That is, the semantic classes and the corresponding entities are constructed first, and then the text and visual contexts corresponding to the entities are collected in turn.

Step 1. Semantic Classes and Entities Collection Wikipedia has compiled a vast list of entities corresponding to semantic classes[1], which are organized in a hierarchical structure. We pick a selection of semantic classes with certain principles (discussed in next Section) and crawl the corresponding entities. In addition, numerous entities randomly sampled from Wikipedia pages are appended to the entity vocabulary as negative entities. Further, polysemous and synonymous entities are also added to the vocabulary as hard negative entities and hard positive entities, respectively.

[1] https://en.wikipedia.org/wiki/List_of_lists_of_lists

Step 2. Entity-Labeled Sentences Collection We crawl Wikipedia articles containing abundant entity mentions with human-annotated hyperlinks[2] that uniquely identify an entity. Since the entities crawled in Step 1 contain hyperlinks, we can utilize these hyperlinks to associate the entities with the respective sentences and convey the textual information to the entities.

Step 3. Related Images Collection In this step, images corresponding to the entities or sentences are acquired through Google Image search engine. To remove the distraction of extraneous content in sentence, keywords in the sentence are extracted with KeyBERT (Grootendorst 2020). We stitch them with the entity name and semantic class as the search query, and obtain the top 10 images of the search results.

Step 4. Images Re-ranking One of the 10 images needs to be selected as the visual information of the entity. An ideal image should reflect the content of the sentence and contain the entity simultaneously. With both aspects in mind, a simple but effective image re-ranking algorithm was devised to select the most appropriate image v_i for sentence t_i and entity e:

$$\begin{aligned} score(v_i, t_i, e) = \alpha\, &\text{CLIP-IMG}(v_i) \odot \text{CLIP-TEXT}(t_i) \\ &+ (1-\alpha) \max_{o_i^j \in \text{Obj}(v_i)} (\text{cos\_sim}(o_i^j, \text{Img}(e))) \end{aligned} \tag{1}$$

The first term measures the relevance of image v_i and sentence t_i, which is what CLIP excels at. The second leveraged FasterRCNN (Ren et al. 2015) to detect objects $\text{Obj}(v_i)$ in image and calculate their similarity to typical image $\text{Img}(e)$ of entity in the Wikipedia Infobox. The second determines whether the entity appears in the image or not. We take the image with the highest score as the one corresponding to the sentence t_i and entity e and leave the exploitation of multiple images for future research.

Human Calibration and Annotation

The dataset automatically generated after the above steps is inevitably noisy. Especially in Steps 3 and 4, A mismatch between images and sentences may exist. To improve the quality of images while verifying the effectiveness of the re-ranking algorithm, we hired human annotators who were required to evaluate the relevance of images to sentences and entities, categorized into three categories: relevant to both (R/T E&S), relevant to only the sentence (R/T S), and irrelevant to both (IR). For images that are irrelevant to both after re-ranking, the annotators need to select a new image.

From Table 1, we observe that the re-ranking algorithm significantly improves the relevance of images to both text and entities, compared to using the Top 1 image returned by the search engine directly. The inter-annotator agreement measured by Fleiss's Kappa (Fleiss 1971) all exceeded 0.8, demonstrating the reliability of the annotation results. The strategy using the Top 1 image has the highest image diversity (measured by the inverse of the average cosine similarity of image embeddings) due to the introduction of substantial irrelevant images. Whereas the first term of the re-ranking algorithm guarantees the relevance of images and sentences while also avoiding a singular selection of typical images of the entity, potentially ensuring that there is no significant decrease in image diversity.

Strategy	R/T E&S (%)	R/T S (%)	IR (%)	Kappa	Diversity
Top 1	52.7	14.8	32.5	0.842	1.813
Re-ranking	78.1	15.2	6.7	0.862	1.792
Annotation	80.8	19.2	0	0.858	1.798

Table 1: Relevance of images between entities and sentences when using different strategies to process images.

Analysis of MESED

MESED is the first multi-modal ESE dataset with meticulous manual calibration. It consists of 14,489 entities collected from Wikipedia, and 434,675 image-sentence pairs. The 70 fine-grained semantic classes in MESED contain an average of 82 entities with a minimum of 23 and a maximum of 362. Each fine-grained class contains 5 queries with three seed entities and 5 queries with five seed entities. MESED may not feature the largest total number of candidate entities, but we believe that the number of entities is not a key factor in measuring the quality of a dataset. Most candidate entities in previous datasets are randomly selected negative entities, which are significantly different from the target entities and do not enhance the challenge of the dataset.

	Wiki	**APR**	**CoNLL**	**ONs**	**MESED**
# Classes	8	3	4	8	70
Granularity	Coarse	Coarse	Coarse	Coarse	Fine
# Queries / Class	5	5	1	1	10
# Seed / Query	3	3	10	10	3/5
# Entities	33K	76K	6K	20K	14K
# Sentences	973K	1043K	21K	144K	434K
Multi-Modal	✗	✗	✗	✗	✓

Table 2: Comparison of ESE datasets.

We ensured that the MESED was challenging from multiple perspectives: (1) We meticulously designed the schema of semantic classes, which consists of three layers of granularity. Fine-grained semantic classes that belong to the same parent class have semantic overlap, making them hard negative semantic classes for each other. (2) We included entities sharing words with the target entities obtained through the BM25-based Wikipedia search engine, as hard negative entities in the candidate word list. (3) We assessed the model's ability to expand synonymous entities by obtaining the entity's synonyms via Wikidata SPARQL and replacing a portion of the entity with synonyms having an edit distance greater than 5 from it. Due to space constraints, more detailed analysis and experiments on MESED are placed **in the appendices in the Supplementary Material, and they are highly recommended to the reader**.

[2] E.g., https://en.wikipedia.org/wiki/Earth

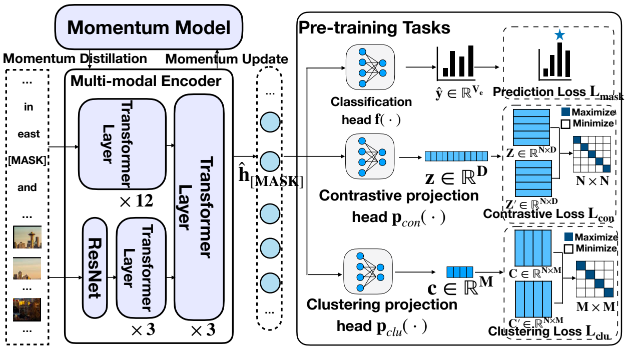

Momentum Model
Momentum Distillation
Momentum Update
Multi-modal Encoder
...
in
east
[MASK]
and
...
...
...
Transformer Layer
× 12
ResNet
Transformer Layer
× 3
Transformer Layer
× 3
$\hat{\mathbf{h}}_{[MASK]}$
...
...
Pre-training Tasks
Classification head $\mathbf{f}(\cdot)$
$\hat{\mathbf{y}} \in \mathbb{R}^{\mathbf{V_e}}$
Prediction Loss $\mathbf{L}_{mask}$
Contrastive projection head $\mathbf{p}_{con}(\cdot)$
$\mathbf{z} \in \mathbb{R}^{\mathbf{D}}$
$\mathbf{Z} \in \mathbb{R}^{\mathbf{N \times D}}$
$\mathbf{Z}' \in \mathbb{R}^{\mathbf{N \times D}}$
Maximize
Minimize
$\mathbf{N \times N}$
Contrastive Loss $\mathbf{L}_{con}$
Clustering projection head $\mathbf{p}_{clu}(\cdot)$
$\mathbf{c} \in \mathbb{R}^{\mathbf{M}}$
$\mathbf{C} \in \mathbb{R}^{\mathbf{N \times M}}$
$\mathbf{C}' \in \mathbb{R}^{\mathbf{N \times M}}$
Maximize
Minimize
$\mathbf{M \times M}$
Clustering Loss $\mathbf{L}_{clu}$

Figure 2: The training framework of the multi-modal entity representation phase.

Methods

Overall Framework

We describe the proposed MultiExpan method for MESE, which expands the initial entity set with multi-modal contexts. Inspired by the previous ProbExpan (Li et al. 2022b), we divide MultiExpan into two steps: multi-modal entity representation phase and entity expansion phase. In the first phase, we design a multi-modal entity-level encoder whose output is the probability distribution of masked span over candidate entities. The entity is represented as the average of the predicted entity distributions for all sentences containing it. Four multi-modal self-learning pre-training tasks are proposed to refine the entity representation. In the second phase, MultiExpan obtains the target entities according to the similarities of the probabilistic representation of the entities. We note that MultiExpan is proposed to provide a robust multi-modal baseline and to explore the effectiveness of different pre-training tasks.

Multi-modal Entity Representation

Multi-modal encoder first processes text and images separately with self-attention Transformer, then combines them for deep cross-modal interaction.

Text Firstly, to handle the text information, we replace entity mentions in sentences with [MASK] to construct the inputs for text modality. Concerning the contextual text $T = \{w_1, w_2, ..., w_{L_1}\}$ with masked entity mention, we directly use 12 layers of Transformer initialized by $\text{BERT}_{\text{BASE}}$ (Kenton and Toutanova 2019) to obtain the textual context's embeddings:

$$\hat{W} = \{\hat{w}_1, \hat{w}_2, ..., \hat{w}_{L_1}\} = \text{BERT}_{\text{BASE}}(T) \tag{2}$$

where L_1 is the max length of tokens in the sentences.

Image Secondly, we deal with the image information. Different from the regional features and grid features widely used in the field of image feature extraction, the patch features we adopt are simple yet efficient. We transform each image into a fixed shape and determine the size of each patch, divide each image into 36 patches $I = \{i_1, i_2, ..., i_{L_2}\}$, and use the backbone Resnet to extract patch features:

$$\{v_1, v_2, ..., v_{L_2}\} = Flat(Resnet(I)) \tag{3}$$

where L_2 is the number of patches and $Flat(\cdot)$ indicates the flatting function that reshapes the patch features extracted from Resnet into one dimensional.

Since the patch features will cause the loss of position information during segmentation, we add a learnable position embedding $P = \{p_1, p_2, ..., p_{L_2}\}$ in order to mark the position information of each patch. Both patch features and position embeddings are combined through pair-wise add.

Finally, we build a 3-layer transformer architecture as image encoder in the visual information processing:

$$\hat{V} = \{\hat{v}_1, \hat{v}_2, ..., \hat{v}_{L_2}\} = Encoder_V(Flat(Resnet(I)) \oplus P) \tag{4}$$

Cross-modal fusion After obtaining the information of the two modalities, the hidden states $\{h_1, h_2, ..., h_L\}$ are obtained through the concatenation of text features and visual features: $concat(\hat{W}, \hat{V})$. Then we feed it into a 3-layer transformer for interaction and fusion between modalities so that the image-text pairs are fully aligned:

$$\{\hat{h}_1, \hat{h}_2, ..., \hat{h}_L\} = Encoder_{cross}(\{h_1, h_2, ..., h_L\}) \tag{5}$$

where $L = L_1 + L_2$ and the structure of the transformer is the same as the above-mentioned visual encoder.

A classification head $\mathbf{f}$ is attached behind the multi-modal encoder. After getting the hidden state of the mask position, the embedding vector is transformed into the probability distribution of the masked entity over the possible candidate entities by MLP and Softmax function:

$$\hat{y} = \mathbf{f}(\hat{h}_{[MASK]}) = Softmax(MLP(\hat{h}_{[MASK]})), \hat{y} \in \mathbb{R}^{V_e} \tag{6}$$

in which V_e is the size of candidate entities vocabulary.

Four self-supervised pre-training objectives are proposed for the training. The multi-modal encoder iteratively optimizes the four objectives:

Masked entity prediction loss With respect to the masked entity prediction task, the model takes images and the masked sentences as input and obtains the entity probability distribution $\hat{y}$ of the masked position as described above. Cross-entropy loss with label smoothing is applied to allow the model to learn the underlying semantics of entities:

$$\begin{aligned}\mathcal{L}_{mask} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{V_e} & y_i[j] \cdot (1-\eta) \cdot log(\hat{y}_i[j]) \\ & +(1-y_i[j]) \cdot \eta \cdot log(1-\hat{y}_i[j])\end{aligned} \tag{7}$$

where the ground truth y is the one-hot vector and N is the batch size. η is the smoothing factor that prevents entities sharing semantics with the target entity from being overly suppressed.

Contrastive learning loss Contrastive learning provides clearer semantic boundaries of semantic classes through drawing the representation of the same semantic class entities closer and the representation of different semantic class entities further apart (Li et al. 2022d,c). We generate the positive and negative entities for each semantic class from the expanded list obtained in the previous iteration. The entities ranked in the top K_{pos} positions are defined as positive entities, while the entities ranked from L_{neg} to U_{neg} are considered negative entities. The samples from positive/negative entities are paired to form positive/negative sample pairs. For a mini-batch of size N, each sample x_{2i-1} forms $2N-1$ pairs with others, among which we pair x_{2i-1}, x_{2i} to be positive and define other $2N-2$ pairs to be negative.

Since directly performing contrastive learning on the hidden features $\hat{h}_{[MASK]}$ may cause information loss, we plugged in a two-layer MLP $\mathbf{p}_{con}(\cdot)$ behind multi-modal encoder to map the hidden features to a normalized subspace via $z_i = \mathbf{p}_{con}(\hat{h}_{[MASK]})$, where $z_i \in \mathbb{R}^D$ and D is the dimension of subspace. The pair-wise similarity is measured by dot product:

$$s(z_i, z_j) = z_i \cdot z_j^\top, i, j \in [1, 2N] \tag{8}$$

The contrastive learning loss that concentrates on hard negative entities is applied. For a given sample z_i (suppose it forms a positive pair with z_j), the loss is defined as:

$$l_i = -\log \frac{e^{s(z_i,z_j)/t}}{e^{s(z_i,z_j)/t} + R_i^-}, \tag{9}$$

$$R_i^- = \max(\frac{-(2N-2)\cdot\tau\cdot e^{s(z_i,z_j)/t} + \widetilde{R_i^-}}{1-\tau^+}, e^{-\frac{1}{t}}) \tag{10}$$

$$\widetilde{R_i^-} = \frac{(2N-2)\sum_{k:k\neq i\neq j} e^{(1+\beta)\cdot s(z_i,z_k)/t}}{\sum_{k:k\neq i\neq j} e^{\beta\cdot s(z_i,z_k)/t}} \tag{11}$$

where τ, β, t are hyperparameters, representing class prior probability, hard negative entity concentration level, and temperature. The contrastive loss in a batch is computed as:

$$\mathcal{L}_{con} = \sum_{i=1}^{2N} l_i \tag{12}$$

Clustering learning loss Similar to contrastive learning, clustering learning attracts positive semantic class pairs and repels negative semantic class pairs. We employ an alternative projection head, denoted as $\mathbf{p}_{clu}$, to map the input sample x_i onto a semantic class subspace, resulting in $c_i = \mathbf{p}_{clu}(\hat{h}[MASK])$. The dimension M of c_i corresponds to the number of clusters, namely the number of target semantic classes. Each element of the feature indicates the probability that it belongs to a particular semantic class. We posit that a semantic class can be characterized by the probabilistic responses of a batch of entities towards it. Formally, let $C = [c_1, \cdots, c_{2i-1}, \cdots, c_{2N-1}] \in \mathbb{R}^{N\times M}$ denotes the class probability distribution under samples $\{x_1, \cdots, x_{2i-1}, \cdots, x_{2N-1}\}$, and $C' = [c_2, \cdots, c_{2i}, \cdots, c_{2N}]$ for samples $\{x_2, \cdots, x_{2i}, \cdots, x_{2N}\}$. The positive clustering pairs are formed by the semantic classes represented by the same columns of matrices C and C', due to the fact that the entities x_{2i-1} and x_{2i}, corresponding to each element of these column vectors, are positive sample pairs originating from the same semantic class. For brevity, we denote the i-th column of C as $\hat{c}_{2i-1}$ and $\hat{c}_{2i}$ for the i-th column of C'. Similarly, dot product is adopted to quantify the similarity between $\hat{c}_i$ and $\hat{c}_j$:

$$\hat{s}(\hat{c}_i, \hat{c}_j) = \hat{c}_i^\top \cdot \hat{c}_j, i, j \in [1, 2M] \tag{13}$$

For each semantic class $\hat{c}_i$, the clustering loss $\hat{l}_i$ is computed in the same way as contrastive loss defined in Equation (9)-(11), which distinguishes $\hat{c}_i$ from other $2M-2$ semantic classes except its positive counterpart $\hat{c}_j$. The clustering loss is finally calculated as:

$$\mathcal{L}_{clu} = \sum_{i=1}^{2M} \hat{l}_i \tag{14}$$

Momentum distillation loss The image-sentence pairs in our MESED are collected from the web, often accompanied by noise, which causes the collected images may be weakly related to the sentences, or the extended entities belonging to the semantic class are not included in ground truth. To alleviate the above problems, we introduce momentum distillation learning. During training, a momentum version of the model is slowly updated by exponentially shifting the momentum factor m: $\theta_\mathrm{t} \leftarrow m\theta_\mathrm{t} + (1-m)\theta_\mathrm{s}$ and the momentum model is used to generate pseudo-labels as additional supervision, preventing the student model overfitting to noise.

The momentum distillation loss is expressed as the KL divergence between the pseudo entities probability distribution $\widetilde{y}$ generated by the momentum model and the predicted $\hat{y}$ of the multi-modal encoder at current iteration:

$$\mathcal{L}_{mod} = -\sum_{i=1}^{m} \widetilde{y_i} log(\widetilde{y_i}) - \widetilde{y_i} log(\hat{y_i})) \tag{15}$$

Entity Expansion

The entity is represented as the average of the predicted entity distributions for all sentences containing it. The semantic class is represented by the weighted average of entities in current expansion set and the weight is dynamically maintained by window search algorithm. In this way, candidate entities with similar distribution are placed in the current set measured by KL divergence.

As the expansion process is not the focus of this work, we use window search and entity re-ranking algorithm from the ProbExpan (Li et al. 2022b) and will not repeat them here.

Experiments

Experiment Setup

Compared Methods We compare three categories of models, the first is the traditional text-based ESE approach, including **SetExpan** (Shen et al. 2017), **CaSE** (Yu et al. 2019), **CGExpan** (Zhang et al. 2020), **ProbExpan** (Li et al. 2022b) and **GPT-3.5**. Of the above models, SetExpan, CaSE are the traditional statistical probability-based approaches, and CGExpan and ProbExpan are the most advanced methods based on pre-trained language model BERT. We also evaluated vision-based models: **VIT** (Dosovitskiy et al. 2020), **BEIT** (Bao et al. 2021) and image encoder of CLIP (**CLIP-IMG**). For multi-modal expansion, we explored multi-modal models with different structures comprising **CLIP** (Radford et al. 2021) and **ALBEF** (Li et al. 2021). Both the above-mentioned vision-based and multi-modal models are further pre-trained via entity prediction tasks, analogous to the method defined in Equation (7).

Modality	**Method**	**‖Seed‖=3**								
		MAP				**P**				**Avg**
		@10	**@20**	**@50**	**@100**	**@10**	**@20**	**@50**	**@100**	
T	SetExpan	26.10	20.98	15.83	13.91	34.25	29.58	24.25	22.96	23.48
	CaSE	27.71	20.93	14.63	12.02	36.85	30.57	24.83	23.63	23.90
	CGExpan	38.89	32.51	24.69	21.06	45.85	39.85	33.19	32.80	33.61
	GPT-3.5	31.10	24.73	19.20	17.07	37.65	31.35	26.08	25.11	26.54
	GPT+Name	42.12	35.32	26.83	23.21	52.32	41.23	35.89	35.73	36.58
	ProbExpan	65.47	57.50	43.96	40.73	71.30	64.35	55.73	51.99	56.38
V	VIT	65.02	55.94	41.89	32.40	67.95	59.53	46.08	36.94	50.72
	BEIT	68.45	58.58	43.59	33.69	71.70	62.13	47.60	37.66	52.93
	CLIP-IMG	66.39	57.04	41.72	32.42	68.85	60.90	45.79	36.81	51.24
T+V	CLIP	76.41	65.75	49.58	40.08	79.20	69.53	53.10	43.66	59.66
	ALBEF	83.55	75.46	63.02	54.47	86.60	79.15	68.03	61.12	71.43
	Ours (MEP)	86.07	79.18	67.66	58.91	89.10	82.85	72.13	65.17	75.13
	Ours (Full)	91.44	86.85	76.86	63.34	93.60	89.63	80.37	67.15	81.16

Modality	**Method**	**‖Seed‖=5**								
		MAP				**P**				**Avg**
		@10	**@20**	**@50**	**@100**	**@10**	**@20**	**@50**	**@100**	
T	SetExpan	25.99	20.64	15.20	13.51	34.90	29.93	24.26	23.29	23.47
	CaSE	32.01	24.63	17.99	14.58	41.50	34.75	28.83	27.03	27.67
	CGExpan	38.86	31.49	23.54	20.23	45.55	38.28	31.88	32.15	32.75
	GPT-3.5	31.79	25.46	20.12	19.94	39.40	33.13	28.67	30.45	28.62
	GPT+Name	42.32	36.48	25.76	22.36	52.94	42.10	34.68	35.12	36.47
	ProbExpan	66.29	59.31	48.90	42.51	73.15	66.78	58.51	54.54	58.75
V	VIT	62.29	55.43	41.30	31.54	68.20	58.93	45.61	35.91	49.90
	BEIT	70.14	59.04	43.08	33.21	73.45	62.93	47.25	37.17	53.28
	CLIP-IMG	67.67	57.28	41.41	31.86	70.40	60.80	45.25	35.94	51.33
T+V	CLIP	77.37	65.92	49.01	39.05	79.80	69.48	52.41	42.50	59.44
	ALBEF	85.04	76.25	62.45	53.64	87.80	79.70	67.37	60.06	71.54
	Ours (MEP)	87.77	79.96	67.24	57.62	90.90	83.55	71.41	63.41	75.23
	Ours (Full)	92.67	87.27	75.70	61.36	94.30	89.68	78.56	64.46	80.50

Table 3: Main experiment results. Text-based, vision-based, and multi-modal expansion methods are evaluated.

Evaluation Metrics The objective of ESE is to expand the ranked entity list based on their similarity to given seed entities in descending order. Following previous research (Zhang et al. 2020; Li et al. 2022b; Yan et al. 2020), two widely used evaluation metrics, MAP@K and P@K, are employed. The MAP@K metric is computed as follows:

$$\text{MAP@}K = \frac{1}{|Q|} \sum_{q \in Q} \text{AP}_K(R_q, G_q) \tag{16}$$

Here, Q is the collection for each query q. AP$_K(R_q, G_q)$ denotes the average precision at position K with the ranked list R_q and ground-truth list G_q. P@K is the precision of the top-K entities. In the experiment, queries with ‖Seed‖=3 and 5 are evaluated separately.

Main Experiment

The results of the main experiment are presented in Table 3, from which we observe that: (1) The multi-modal methods outperform the mono-modal methods in general. Remarkably, our MultiExpan achieves superior performance solely by employing masked entity prediction (MEP) task. Moreover, the full version of MultiExpan achieves the best performance.

(2) In terms of the structure of multi-modal models, ALBEF and our MultiExpan exhibit deep modality interaction through the Transformer, which is better suited for the ESE task compared to the CLIP's shallow modal interaction via dot product similarity calculation. These results indicate that deep modal interaction and fusion is a direction that can be explored in the future.

(3) In terms of the vision-based models, BEIT excels in leveraging finer-grained image semantics, such as object and

background information, by pre-training on masked image modeling. In contrast to the VIT model which learns the overall image semantics through image classifying images in the Image Net dataset, BEIT demonstrates better results in entity understanding. Meanwhile, the image encoder of CLIP also captures richer semantics than the VIT model owing to its linkage with the text modality. However, relying solely on image modality does not suffice to produce satisfactory results, and the text modality still remains dominant.

(4) The increase of ||Seed|| does not necessarily translate to an enhancement in overall performance. More seeds can describe the semantic classes more precisely and retrieve some "must be correct" entities more safely, so MAP/P improves when K is small (=10,20). However, more seed entities mean a larger search space for semantic classes, necessitating a more meticulous analysis of common entity properties than the current model allows. This issue represents the persistent challenge of semantic drift that confronts ESE models, so MAP/P decreases when K is larger. Of course, increasing ||Seed|| helps disambiguate the query with entities belonging to multiple classes. Such as in the semantic class *Light Novel*, where some seed entities also are *Manga*, increasing ||Seed|| makes a gain of 17.5% average on all metrics.

(5) GPT-3.5 did not achieve satisfactory results, and was even inferior to unsupervised CGExpan. Through meticulous examination of GPT-3.5's performance on specific semantic classes, we discovered that model struggled with complex classes (e.g., *108 Martyrs of World War II*). We explicitly instructed GPT-3.5 to reason about the class names first, and then expand based on them. This modification, named GPT+Name, exhibited a substantial improvement. This approach aligns with the idea of emerging chain-of-thought reasoning (Wei et al. 2022) for large language models (Touvron et al. 2023; Li et al. 2023c; Yu et al. 2023b), i.e., thinking step by step. We suggest future research to explore the combination of chain-of-thought and ESE tasks.

Pre-training Tasks Analysis

We compared the effects of different pre-training tasks on MultiExpan. The masked entity prediction task enables the model to learn the underlying semantics of entities, which is further enhanced by the addition of three pre-training tasks. The results presented in Table 4 demonstrate that each pre-training task confers a gainful effect on the model. Notably, we found that contrastive learning with hard negative entities yields the greatest performance improvement for the model, by providing clearer semantic boundaries. While clustering learning brings comparable gains to contrastive learning at MAP/P@K=10 and 20, it is less effective at larger K. This is because contrastive learning operates directly on entities and more directly aggregates target entities into tight clusters. In contrast, momentum distillation learning brings a smaller performance gain, which we believe is mainly attributed to its ability to prevent overfitting in the presence of noisy data. This observation underscores the high quality of the data provided by MESED, particularly the accurate annotation of entities in sentences.

Extensive experiments on the hyperparameters sensitivity of the pre-training tasks are presented in Appendix, demonstrating the robustness of MultiExpan to the parameters.

Modality Analysis

We also carry out analysis experiments on each modality to answer the following questions.

Are the multiple modalities complementary?

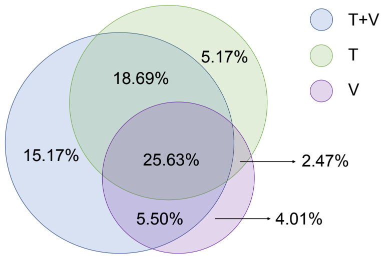


Figure 3: The contribution of each modality.

We present a Venn diagram illustrating the impact of different modalities on MESE, as depicted in Figure 3. T, V and T+V represent ProbExpan, BEIT and our MultiExpan respectively. The size of each circle corresponds to the proportion of the top 100 ranked entities that belong to the ground truth, and the intersection of the circles represents the overlap of entities. Our analysis shows that the textual modality still prevails over the visual modality. Whereas the visual modality is introduced as supplementary information, 15.17% of the target entities in MultiExpan are sorted to a higher position, while 5.17% of the entities that were originally correctly expanded are excluded, due to the image noise.

Is it better to have multi-modal contexts of both seed and candidate entities? During the inference phase, we separately removed the textual and visual information from the candidate or seed entities in MultiExpan. The resulting performances are shown in the last 6 rows of Table 5, with subscripts indicating the operations performed on seeds (s) or candidates (c). Our results indicate that removing any part of the modal information for any part of the entities is detrimental to the overall performance. However, when particular modal information was removed from seed entities, it caused severe performance degradation, whereas removing modal information from candidate entities caused only a slight performance loss. These findings suggest that modeling the semantics of the seed entity set is more crucial than modeling individual entities. Additionally, MultiExpan demonstrated a decrease in performance when we removed the input text or images during the pre-training phase, further demonstrating its ability to effectively utilize multi-modal information.

What visual clues are provided by the visual modality? We randomly sample 200 entities and determine that images can provide essential visual clues, including (1) Objects, which can augment the limited textual information by depicting the entities themselves, (2) Scenes, which showcase the environment where the entity exists to differentiate between the target semantic class and the hard negative semantic class, e.g., indoor vs. outdoor, water vs. land, (3) Properties, which demonstrate the common traits of entities to align entities of the same class, such as appearance of *Cats*, and (4) Other:

Model	MAP				P				Avg
	@10	@20	@50	@100	@10	@20	@50	@100	
MultiExpan (MEP)	86.07	79.18	67.66	58.91	89.10	82.85	72.13	65.17	75.13
+ Contrastive	90.71	86.58	75.58	62.69	93.35	89.60	79.23	67.10	80.61
+ Clustering	89.10	82.83	70.85	60.48	91.65	86.05	74.75	65.92	77.70
+ Distillation	86.97	80.48	68.30	59.43	89.85	83.65	72.34	65.23	75.78
MultiExpan (Full)	91.44	86.85	76.86	63.34	93.60	89.63	80.37	67.15	81.16

Table 4: Comparison of different pre-training tasks.

Model	MAP				P				Avg
	@10	@20	@50	@100	@10	@20	@50	@100	
MultiExpan (MEP)	86.07	79.18	67.66	58.91	89.10	82.85	72.13	65.17	75.13
pre-train w/o T	65.97	57.87	42.84	33.39	70.45	62.50	48.85	39.70	52.70
pre-train w/o V	66.87	60.18	52.26	47.57	73.90	68.13	62.45	60.88	61.53
w/o T_s and T_c	20.67	18.32	13.13	9.66	27.80	26.10	21.54	18.17	19.42
w/o T_s	20.75	18.43	13.57	9.92	27.50	25.88	21.54	18.34	19.49
w/o T_c	85.45	77.99	66.53	56.58	88.10	81.95	71.17	62.68	73.81
w/o V_s and V_c	58.99	50.36	40.38	35.60	64.05	56.53	48.37	47.09	50.17
w/o V_s	60.44	51.95	41.92	37.18	65.05	57.55	49.25	47.67	51.38
w/o V_c	84.79	76.94	64.55	55.92	87.90	81.18	69.76	63.07	73.01

Table 5: Ablation study on modality absence.

Visual Clues	Proportion	P@100	
		ProbExpan	MultiExpan
Object	46.3%	57.44	70.21
Scene	21.2%	67.44	72.09
Property	22.2%	66.66	80.00
Others	3.4%	61.90	76.19

Table 6: Model performance under different visual clues.

Other important visual clues. We annotate 200 entity images with their corresponding visual clue types and assess MultiExpan's capacity to leverage different visual clues. As Table 6 shows, all types of visual cues are beneficial to MESE, and visual modalities mainly supplement the textual information by highlighting objects in the images. In contrast, MultiExpan utilizes scenes to a lesser extent as they represent more abstract concepts.

Case studies, visual clues examples and detailed performance on each semantic class can be found in Appendix.

Conclusion

In this paper, we introduce a novel task called Multi-modal Entity Set Expansion (MESE), which aims to leverage multiple modalities to represent and expand entities. The MESED dataset is the first multi-modal dataset for ESE with fine-grained semantic classes and hard negative entities. In addition, A powerful multi-modal model MultiExpan is proposed which is pre-trained on four multimodal pre-training tasks. MultiExpan achieves state-of-the-art results compared to other mono/multi-modal models. In the future, we will investigate the applicability of generative PLMs, such as GPT-4, in addressing MESE task. MESED can also serve as a reliable benchmark for assessing the multi-modal entity understanding capacities of large PLMs.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), the Major Key Project of PCL for Experiments and Applications (PCL2021A06), Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (HW2021008), and Shenzhen Science and Technology Program (WDZC20231128091437002).

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Chen, Z.; Cafarella, M.; and Jagadish, H. V. 2016. Long-Tail Vocabulary Dictionary Extraction from the Web. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, 625–634. New York, NY, USA: Association for Computing Machinery. ISBN 9781450337168.

Cheng, X.; Cao, B.; Ye, Q.; Zhu, Z.; Li, H.; and Zou, Y. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6492–6505.

Cheng, X.; Dong, Q.; Yue, F.; Ko, T.; Wang, M.; and Zou, Y. 2023b. M 3 st: Mix at three levels for speech translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Cheng, X.; Xu, W.; Zhu, Z.; Li, H.; and Zou, Y. 2023c. Towards spoken language understanding via multi-level multi-grained contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 326–336.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Grootendorst, M. 2020. KeyBERT: Minimal keyword extraction with BERT.

Henriksson, A.; Moen, H.; Skeppstedt, M.; Daudaravičius, V.; and Duneld, M. 2014. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of biomedical semantics*, 5(1): 1–25.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Lauscher, A.; Vulić, I.; Ponti, E. M.; Korhonen, A.; and Glavaš, G. 2020. Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1371–1383.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, Y.; Chen, J.; Li, Y.; Xiang, Y.; Chen, X.; and Zheng, H.-T. 2023a. Vision, Deduction and Alignment: An Empirical Study on Multi-Modal Knowledge Graph Alignment. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Li, Y.; Huang, S.; Zhang, X.; Zhou, Q.; Li, Y.; Liu, R.; Cao, Y.; Zheng, H.; and Shen, Y. 2022a. Automatic Context Pattern Generation for Entity Set Expansion. *CoRR*, abs/2207.08087.

Li, Y.; Li, Y.; Chen, X.; Zheng, H.-T.; and Shen, Y. 2023b. Active relation discovery: Towards general and label-aware open relation extraction. *Knowledge-Based Systems*, 282: 111094.

Li, Y.; Li, Y.; He, Y.; Yu, T.; Shen, Y.; and Zheng, H.-T. 2022b. Contrastive Learning with Hard Negative Entities for Entity Set Expansion. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1077–1086.

Li, Y.; Ma, S.; Wang, X.; Huang, S.; Jiang, C.; Zheng, H.-T.; Xie, P.; Huang, F.; and Jiang, Y. 2023c. EcomGPT: Instruction-tuning Large Language Model with Chain-of-Task Tasks for E-commerce. *arXiv preprint arXiv:2308.06966*.

Li, Y.; Ma, S.; Zhou, Q.; Li, Z.; Yangning, L.; Huang, S.; Liu, R.; Li, C.; Cao, Y.; and Zheng, H. 2022c. Learning from the Dictionary: Heterogeneous Knowledge Guided Fine-tuning for Chinese Spell Checking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 238–249. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Li, Y.; Zhou, Q.; Li, Y.; Li, Z.; Liu, R.; Sun, R.; Wang, Z.; Li, C.; Cao, Y.; and Zheng, H.-T. 2022d. The Past Mistake is the Future Wisdom: Error-driven Contrastive Probability Optimization for Chinese Spell Checking. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3202–3213. Dublin, Ireland: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*.

Schumacher, E.; and Dredze, M. 2019. Learning unsupervised contextual representations for medical synonym discovery. *JAMIA open*, 2(4): 538–546.

Shen, J.; Wu, Z.; Lei, D.; Shang, J.; Ren, X.; and Han, J. 2017. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 288–304. Springer.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E. H.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Yan, L.; Han, X.; He, B.; and Sun, L. 2020. Global bootstrapping neural network for entity set expansion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3705–3714.

Yu, P.; Huang, Z.; Rahimi, R.; and Allan, J. 2019. Corpus-based set expansion with lexical features and distributed representations. In *Proceedings of the 42nd International*

ACM SIGIR Conference on Research and Development in Information Retrieval, 1153–1156.

Yu, T.; Hu, J.; Yao, Y.; Zhang, H.; Zhao, Y.; Wang, C.; Wang, S.; Pan, Y.; Xue, J.; Li, D.; Liu, Z.; Zheng, H.-T.; and Sun, M. 2023a. Reformulating Vision-Language Foundation Models and Datasets Towards Universal Multimodal Assistants. arXiv:2310.00653.

Yu, T.; Jiang, C.; Lou, C.; Huang, S.; Wang, X.; Liu, W.; Cai, J.; Li, Y.; Li, Y.; Tu, K.; Zheng, H.-T.; Zhang, N.; Xie, P.; Huang, F.; and Jiang, Y. 2023b. SeqGPT: An Out-of-the-box Large Language Model for Open Domain Sequence Understanding. arXiv:2308.10529.

Yu, T.; Li, Y.; Chen, J.; Li, Y.; Zheng, H.-T.; Chen, X.; Liu, Q.; Liu, W.; Huang, D.; Wu, B.; and Wang, Y. 2023c. Knowledge-augmented Few-shot Visual Relation Detection. arXiv:2303.05342.

Zhang, Y.; Shen, J.; Shang, J.; and Han, J. 2020. Empower Entity Set Expansion via Language Model Probing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8151–8160.