SemEval-2023 Task 1: Visual Word Sense Disambiguation

Alessandro Raganato[♡] **Iacer Calixto**[◇] **Asahi Ushio**[♣]
Jose Camacho-Collados[♣] **Mohammad Taher Pilehvar**[♠]
[♡] DISCo, University of Milano-Bicocca, Italy
[◇] Amsterdam UMC, University of Amsterdam, Netherlands
[♣] School of Computer Science and Informatics, Cardiff University, United Kingdom
[♠] Tehran Institute for Advanced Studies, Khatam University, Iran
[♡]alessandro.raganato@unimib.it, [◇]i.coimbra@amsterdamumc.nl,
[♣]ushioa@cardiff.ac.uk, [♣]camachocolladosj@cardiff.ac.uk,
[♠]mp792@cam.ac.uk

Abstract

This paper presents the Visual Word Sense Disambiguation (Visual-WSD) task. The objective of Visual-WSD is to identify among a set of ten images the one that corresponds to the intended meaning of a given ambiguous word which is accompanied with minimal context. The task provides datasets for three different languages: English, Italian, and Farsi. We received a total of 96 different submissions. Out of these, 40 systems outperformed a strong zero-shot CLIP-based baseline (Radford et al., 2021). Participating systems proposed different zero- and few-shot approaches, often involving generative models and data augmentation. More information can be found on the task's website: https://raganato.github.io/vwsd/.

1 Introduction

Word Sense Disambiguation (WSD) is the task of associating a word in context with its intended sense, generally from a pre-defined sense inventory. While there has been significant progress in the last few years (Bevilacqua et al., 2021; Loureiro et al., 2021), mainly powered by progress in language models, WSD has been mainly limited to settings with textual content only. In most settings, WSD uses sense inventories obtained from lexical resources, such as WordNet (Miller, 1998). However, in real-world scenarios, WSD is often associated with other modalities, such as images. The Visual Word Sense Disambiguation (Visual-WSD) task aims at filling this gap: given a word and some limited textual context (often a single word), the task is to select among a set of candidate images the one which corresponds to the intended meaning of the target word. Figure 1 provides a simplified overview of the task.

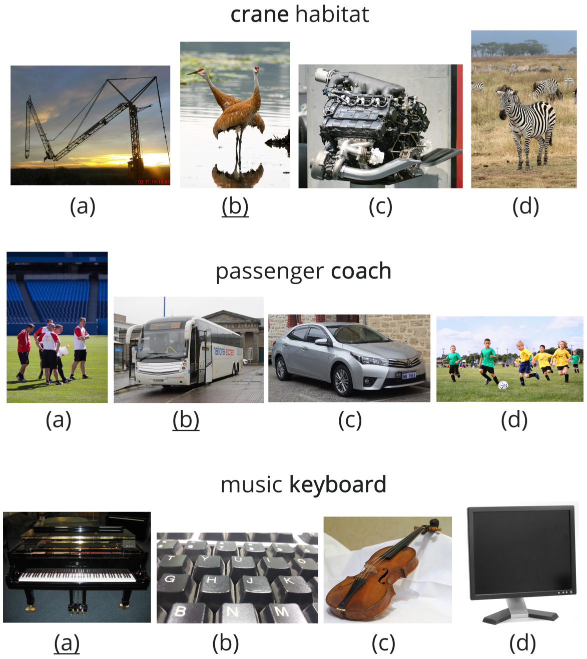

Figure 1: Three examples for the Visual-WSD task. The target (ambiguous) word (in bold) is provided with minimal context (one or two words). The task is to associate the intended sense with the relevant image (underlined). Note: For the task, nine negative images were provided, but for the sake of simplicity, we only show three negative images.

To make the task more challenging, the dataset is constructed with the following objectives in mind: (1) at least one image associated with an incorrect meaning of the word is listed among the options; (2) minimal context (one or two words) is provided for the target word to increase the disambiguation difficulty; and (3) other negative examples are sampled from the same domain as the target sense. Objective (1) guarantees that the task pinpoints systems' abilities in not only identifying the correct class, but also in doing that at the more fine-grained word sense level. Objective (2) ensures a challenging disambiguation setting in which the system is provided with a very limited set of contextual triggers

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2227–2234
July 13-14, 2023 ©2023 Association for Computational Linguistics

to identify the intended meaning. This is relevant given the well-known disambiguation issues for image recognition models, as also observed for recent image generation systems, such as DALLE-2 (Ramesh et al., 2022; Rassin et al., 2022). In objective (3), we build a 'challenge set' where we source images from the same domain as the domain of the target sense, and include a large number of these incorrect images for disambiguation.

The Visual-WSD task can be viewed from two different perspectives. The task is a fine-grained image classification task (in a constrained setting) where the model needs to have knowledge about the different meanings of an ambiguous word. Alternatively, Visual-WSD is a novel disambiguation task where the sense distinctions are given in the visual modality rather than the conventional textual definitions obtained from lexical sense inventories (Miller et al., 1990; Navigli and Ponzetto, 2012).

Task formulation. Given a target ambiguous word (e.g., *coach*) and a limited context (e.g. *passenger*) along with ten images, the Visual-WSD task consists of identifying the most appropriate image for the intended meaning of the ambiguous word. Figure 1 shows a simplified summary of the task with three examples.

Applications. Multimodality is highly intertwined with different aspects of our day-to-day lives. In fact, retrieving and understanding text with textual data is relevant for a myriad of tasks, from object detection to image retrieval (Calixto and Liu, 2017; Gella et al., 2017; Li et al., 2020). Our task enables the in-depth investigation of one aspect often neglected, which is the inherent ambiguity in language (Calabrese et al., 2020a).

Multilinguality. In addition to English, we provide labeled data (context words for given ambiguous targets) for Italian and Farsi. This data serves to analyse the performance of models in languages other than English, and in practical cross-lingual settings where initial training data might not be available for a given language.

2 Data

In this section we detail the data collection and annotation process carried out for the training and testing data. We then provide some statistics on the datasets.

2.1 Data sources

The data for this task is mostly obtained from Wikidata,[1] OmegaWiki,[2] and BabelPic (Calabrese et al., 2020b). Wikidata and OmegaWiki are both collaborative projects to produce a free and open knowledge base and dictionary respectively. Each article, in both resources, can be considered as a concept (or named entity) for which one or multiple related images are provided. BabelPic provides images associated to BabelNet (Navigli and Ponzetto, 2012), a large multilingual encyclopedic dictionary, about non-concrete or abstract concepts.[3] We use BabelNet as a bridge to link the three resources: Wikidata, OmegaWiki and BabelPic. In the following section we describe the semi-automatic process to construct the Visual-WSD dataset.

2.2 Construction procedure

Each instance in the dataset consists of a target word in the context of one or two trigger words, associated with 10 different images: one image corresponding to the intended meaning, and the others referring either to the other meanings of the ambiguous target word, to similar words from the same domain, or other randomly-selected concepts from the used resources.

Training data. We provide silver training data in English by leveraging BabelNet semantic network structure. Specifically, we first collect a list of senses, either ambiguous or monosemous, belonging to the WordNet portion of BabelNet, together with their associated picture provided by BabelNet. Context words are provided based on each concept hypernym (e.g., chef cook as the context for the word *chef*). Moreover, to avoid human faces as potential target images, we also filter out senses denoting people through the associated WordNet category.[4] This training data is intended for silver data only.

Testing data. We provide gold testing data in English, Farsi and Italian. In detail, we first collect a list of ambiguous word senses (the potential target words) with their respective images and definitions from BabelNet. Then, for each word sense, we ask an annotator to provide one or two trigger words (the context) that are enough to identify the

[1] https://www.wikidata.org/

[2] http://www.omegawiki.org/

[3] For our task, we use the gold version only.

[4] https://wordnet.princeton.edu/documentation/lexnames5wn

	Trial	Training	Test
English	16	12869	463
Farsi	-	-	200
Italian	-	-	305

Table 1: Number of instances for trial, training and test.

intended meaning of the word sense when considering the association between definition and image. These trigger words were also selected to be challenging enough as not to give away the meaning of the image in isolation (i.e., the target word is generally necessary to understand the full context). This step is to ensure a challenging text disambiguation task. In the case of English, annotators were proficient English speakers, including the authors of this paper, while in the case of Farsi and Italian, native speakers performed the task. To build the negative sample images, we pick them in a manner to ensure we mitigate model bias to unwanted dataset artifacts. To this end, we construct an *extension* set containing both a set of random images and images that correspond to words other than the target that belong to the same domain (e.g., in Figure 1 we add “zebra” for the target word “crane”). These domains, referred to as BabelDomains (Camacho-Collados and Navigli, 2017), extend the Wikipedia featured article page domains[5] to cover most WordNet and Wikipedia. Finally, for each previously validated instance, a human annotator selects nine images from the *extension* set of candidate images, fulfilling the criteria of selecting related images but different from the target one.

2.3 Statistics

Table 1 shows the number of instances in each split, trial, training, and test. The trial data provided with a few gold examples only, was used for early development purposes during the initial phase of the shared task. Similarly to the test set, trial instances were validated by annotators. English is the largest dataset, including more than 12 thousand silver training instances, and 463 gold test annotations. Concerning Farsi and Italian languages, we provide gold testing data only, with 200 and 305 annotations respectively.

3 Evaluation

We first introduce the metrics that are used to evaluate the participating systems (Section 3.1). Then, we provide a brief description of the baseline system (Section 3.2) as well as the participting systems (Section 3.3), and present the overall results of the task (Section 3.4).

3.1 Evaluation metrics

We use mean reciprocal rank (MRR) and hit rate at 1 (HIT@1) as the evaluation metrics. Given $r = [r_1, \ldots, r_n]$ as the image ranking predictions provided by a given system, MRR is defined as:

$$\mathrm{MRR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i} \tag{1}$$

where n is the number of images. HIT@1 is the ratio of instances ranking the true image as first, that is defined as:

$$\mathrm{HIT@1} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_1(r_i) \tag{2}$$

where $\mathbb{1}_1(r)$ is an indicator function that returns 1 if r is 1, and otherwise 0. HIT@1 can also be viewed as accuracy when only one prediction is provided.

3.2 Baselines

As a baseline for the Visual-WSD task, we use CLIP (Radford et al., 2021), a recent language-vision multi-modal embedding model. We compare the CLIP embedding of the query phrase with the CLIP embedding of each candidate image, and the candidate image with the highest cosine similarity to the query is considered as the prediction. We use the original CLIP for English and multilingual CLIP released by Sentence Transformers (Reimers and Gurevych, 2019) for non-English, where the model's weights are shared via HuggingFace (Wolf et al., 2020).[6]

3.3 Participating systems

Most of our participants' submissions used pre-trained vision-and-language models in a zero-shot

[5] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

[6] Specifically, we used the openai/clip-vit-large-patch14-336 model for CLIP and sentence-transformers/clip-ViT-B-32-multilingual-v1 for multilingual CLIP. The code to reproduce the baseline can be found at https://github.com/asahi417/Visual-WSD-baseline.

Team Name (Run)	HIT@1	MRR
TAM of SCNU (run#2) (Yang et al., 2023)	**72.56**	**82.22**
SRC - Beijing (run#2) (Zhang et al., 2023b)	71.83	80.72
zywiolak (run#2) (Dadas, 2023)	70.49	79.80
Rahul (run#1) (Patil et al., 2023)	69.81	78.23
Chicky (run#2)	68.51	78.80
tara101 (run#2)	62.36	74.20
ResearchTeam_HCN (run#1)	61.82	72.13
ML Mob (run#1) (Poth et al., 2023)	58.94	71.43
calpt (run#1)	58.05	71.27
arshandalili (run#2) (Ghahroodi et al., 2023)	57.46	71.13
ECNU_MIV (run#2) (Li et al., 2023c)	56.43	70.10
QiZhang (run#2)	54.85	68.65
PolitoTeam (run#1) (Vaiani et al., 2023)	53.88	68.13
mmdreza.molavi (run#2) (Molavi and Zeinali, 2023)	53.50	67.92
begab (run#1) (Berend, 2023)	52.96	68.17
PDS2022/23 (run#1)	48.97	63.70
UAlberta (run#1) (Ogezi et al., 2023)	48.41	65.17
CMC MSU (run#1)	47.83	63.76
xiaohuaaa (run#2)	47.38	61.88
floschne (run#1) (Schneider and Biemann, 2023)	46.49	62.19
lky199606 (run#2) (Li et al., 2023b)	44.87	60.81
xiaotian (run#1)	43.00	50.50
Silvilla (run#1)	40.75	48.31
abaaba101 (run#1)	40.27	48.14
Baseline organizers (CLIP)	37.20	54.39
PMCoders (run#2) (Pirhadi et al., 2023)	36.64	54.46
omid (run#2)	35.59	52.18
teamPN (run#2) (Katyal et al., 2023)	32.54	48.97
yjs (run#1)	30.56	48.32
zsbf (run#2)	30.42	41.58
StFX NLP (run#2) (Wei and King, 2023)	29.39	46.29
liliqiqi (run#2)	28.48	46.41
keyi_li (run#2)	28.30	39.41
newtonysls (run#1)	26.60	45.79
RCLN (run#2) (Mijatovic et al., 2023)	22.68	35.02

Table 2: HIT@1 and MRR averaged over all the languages, where only the best run is displayed from each unique user. Best result in each metric is shown in bold.

Team Name (Run)	HIT@1	MRR
Samsung Research China (SRC) - Beijing (run#2)	**84.02**	**89.56**
Rahul (run#1)	83.15	88.80
yixuan_qiao (run#2)	81.86	87.08
TAM of SCNU (run#2)	80.13	87.42
chrisz (run#2)	78.83	85.87
zywiolak (run#2)	77.97	85.88
Chicky (run#1)	74.08	83.80
xiaohuaaa (run#1)	74.08	83.49
tara101 (run#2)	74.08	83.49
ResearchTeam_HCN (run#2)	72.35	80.80
ML Mob (run#1)	71.06	82.22
calpt (run#1)	71.06	82.18
mgrebowiec (run#1) (Grębowiec, 2023)	70.84	81.70
omid (run#2)	69.76	80.51
arshandalili (run#1)	69.76	80.51
shan95 (run#1)	68.47	79.17
GLP (run#1) (Zhang et al., 2023a)	68.47	79.17
begab (run#1)	67.82	80.00
teamPN (run#2)	66.95	78.64
CMC MSU (run#2)	66.31	78.55
Straw hat & Mustache (run#1)	65.87	78.22
QiZhang (run#2)	65.66	78.08
PMCoders (run#2)	65.44	78.01
PolitoTeam (run#1)	65.23	77.81
PDS2022/23 (run#1)	65.23	77.46
WeiJinroad (run#1)	64.58	76.97
floschne (run#2)	64.36	77.52
keyi_li (run#1)	63.93	76.57
lky199606 (run#2)	63.93	76.57
ECNU_MIV (run#2)	62.20	75.53
rishabhgarodia (run#1)	60.91	74.59
Baseline organizers (CLIP)	60.48	73.88
HU (run#2) (Diem et al., 2023)	59.61	73.81
jiesli (run#1) (Li et al., 2023a)	59.18	73.21
StFX NLP (run#1)	59.18	73.01
Anderson (run#2)	58.96	73.45
liliqiqi (run#1)	57.45	71.83
mmdreza.molavi (run#1)	57.24	72.05
HHU (run#1)	56.80	72.22
Ebham (run#1) (Taghavi et al., 2023)	56.80	71.68
UAlberta (run#1)	56.80	71.75
zsbf (run#1)	53.56	69.08
yjs (run#1)	47.73	64.20
RCLN (run#1)	43.41	62.48
newtonysls (run#1)	32.18	52.62
stefy_rzv (run#1)	26.78	46.75
UoR-NCL (run#2) (Markchom et al., 2023)	20.52	41.48

Table 3: HIT@1 and MRR on English test set, where only the best run is displayed for each unique user.

setting, similarly to our baseline (i.e., CLIP). A common approach across submissions involved using *generative models* and *data augmentation*. For text inputs, participants used pretrained language models to generate sentences that include the context words, and back-translation (Sennrich et al., 2016) for data augmentation. In the multimodal setting, participants used pretrained text-to-image models (Rombach et al., 2021) to generate candidate images, and used measures of distance between the generated images and the candidate images as a way to find the correct image for some text input. Finally, participants have also used lexical knowledge graphs such as WordNet to enrich input text with sense information, as well as external resources such as Wikipedia as a source from which to retrieve sentences.

In the following, we describe the top four performing systems in more detail, which include the top-performing models for each individual language.

TAM of SCNU (Yang et al., 2023) This system is based on a Fine-grained Contrastive Language-Image Learning (FCLL) model that learns fine-grained image-text knowledge through a new fine-grained contrastive learning mechanism. The contextual information is enriched by establishing a relationship between concepts and sentences. Finally, this model benefits from a newly constructed multilingual and multimodal knowledge base for ambiguous words. This system achieved the top overall result in the task with an average 72.56 accuracy (or Hits@1) performance, which highlights its robustness across languages.

Team Name (Run)	HIT@1	MRR
zywiolak (run#2)	**64.0**	**74.39**
TAM of SCNU (run#2)	60.5	73.19
Samsung Research China (SRC) - Beijing (run#2)	59.0	70.51
Chicky (run#2)	59.0	70.51
xiaotian (run#1)	58.5	70.50
abaaba101 (run#1)	58.5	69.91
Silvilla (run#1)	57.0	67.90
ECNU_MIV (run#2)	53.0	65.80
tara101 (run#2)	53.0	65.80
arshandalili (run#2)	49.5	64.23
mmdreza.molavi (run#2)	48.5	63.14
ResearchTeam_HCN (run#1)	43.0	55.78
QiZhang (run#2)	42.5	57.19
Rahul (run#1)	42.0	56.84
PDS2022/23 (run#1)	42.0	56.58
ML Mob (run#1)	41.5	56.23
PolitoTeam (run#1)	41.0	56.64
calpt (run#1)	38.5	55.41
CMC MSU (run#1)	38.0	55.54
begab (run#1)	35.0	53.47
UAlberta (run#1)	34.0	53.07
Baseline organizers (CLIP)	28.5	46.70
newtonysls (run#1)	24.0	43.04
floschne (run#1)	23.5	42.77
lky199606 (run#2)	21.5	41.79
xiaohuaaa (run#2)	21.5	38.61
yjs (run#1)	21.0	38.99
omid (run#2)	17.0	35.92
PMCoders (run#2)	13.0	34.94

Table 4: HIT@1 and MRR on Farsi test set, where only the best run is displayed from each unique user.

Team Name (Run)	HIT@1	MRR
Rahul (run#1)	**84.26**	**89.05**
jp854 (run#2)	80.00	84.40
TAM of SCNU (run#2)	77.05	86.05
ResearchTeam_HCN (run#1)	73.77	81.87
Chicky (run#2)	72.46	82.08
Samsung Research China (SRC) - Beijing (run#2)	72.46	82.08
xiaotian (run#1)	70.49	80.98
zywiolak (run#2)	69.51	79.15
Silvilla (run#1)	65.25	77.02
calpt (run#1)	64.59	76.21
ML Mob (run#1)	64.26	75.85
abaaba101 (run#1)	62.30	74.51
xiaohuaaa (run#1)	60.00	73.31
tara101 (run#2)	60.00	73.31
PolitoTeam (run#2)	56.72	70.77
QiZhang (run#2)	56.39	70.68
begab (run#1)	56.07	71.05
mmdreza.molavi (run#2)	54.75	68.56
UAlberta (run#1)	54.43	70.69
ECNU_MIV (run#2)	54.10	68.96
arshandalili (run#2)	53.11	68.65
floschne (run#1)	52.46	67.30
lky199606 (run#2)	49.18	64.09
CMC MSU (run#1)	44.59	60.33
PDS2022/23 (run#1)	39.67	57.07
zsbf (run#2)	37.70	55.66
PMCoders (run#2)	31.48	50.43
RCLN (run#2)	26.56	46.00
newtonysls (run#1)	23.61	41.69
yjs (run#1)	22.95	41.79
Baseline organizers (CLIP)	22.62	42.61
StFX NLP (run#1)	21.97	41.28
keyi_li (run#2)	20.98	41.66
liliqiqi (run#2)	20.00	39.81
omid (run#2)	20.00	40.10
teamPN (run#2)	19.67	37.99

Table 5: HIT@1 and MRR on Italian test set, where only the best run is displayed from each unique user.

Samsung Research China - Beijing (Zhang et al., 2023b) In this model, definitions and synonyms of the target word are collected from WordNet, BabelNet, Wikipedia and `vocabulary.com` to build the reference sense inventory. Moreover, images of phrases are collected from the LAION open dataset (Schuhmann et al., 2021) using CLIP retrieval (Cherti et al., 2022). Then, the most suitable definition from the sense inventory is selected using a biencoder architecture with SimCSE (Gao et al., 2021) as the backbone. The matching model is a large version of CLIP trained on LAION-2B from Open CLIP. For Farsi and Italian, senses are directly translated from English. This system achieved the overall second best performance and the best result by a substantial margin in the English test set.

zywiolak (Dadas, 2023) This hybrid system combines multimodal embeddings and knowledge-based approaches. The main classifier is based on the CLIP model, whose results are enriched with additional information retrieved from Wikipedia and lexical databases. The various modules of the system are integrated using a learning to rank (LTR) model. This model takes as input a feature vector describing the results of the individual components of the system and outputs a relevance ranking of candidate images. This system ranked third overall, while being the top performing system in Farsi.

Rahul (Patil et al., 2023) This system presents an ensemble of different neural models. First, CLIP models are used for English, with text-to-text translation models for Farsi-to-English and Italian-to-English. Additionally, this system integrates multilingual BERT-base embeddings (Devlin et al., 2019) for text and ResNet101 embeddings (He et al., 2016) for the image. This system ranked fourth overall in terms of accuracy (Hits@1) and first in the Italian test set.

3.4 Results

Table 2 shows the overall results (averaged over the three languages, English, Italian and Farsi) of all participating systems. Across the board, systems in the top block, i.e. top five submissions, achieve better results than all the others. Tables 3, 4, 5 show language-specific performances for each submission (English, Farsi and Italian, respectively). As can be seen, although top systems for English and Italian achieve similar results, their performance appears to be relatively weaker in managing Farsi data.

4 Conclusions

In this paper, we presented the first task on Visual Word Sense Disambiguation. Instead of the usual single-modality tasks of textual word sense disambiguation and image recognition, our proposal merges these two paradigms into a single unified task. In particular, WSD is reframed by replacing the usual sense inventory based on lexical resources by a dynamic inventory based on images. The task received 96 submissions. The evaluation showed promising results with 40 submissions outperforming a very competitive image recognition baseline based on a zero-shot CLIP model for different languages. Nonetheless, given the recency and challenging nature of the task, there is clear room for improvement for future work.

References

Gábor Berend. 2023. SzegedAI at SemEval-2023 Task 1: Applying Quasi-Symbolic Representations in Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020a. Evilbert: Learning task-agnostic multimodal sense embeddings. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 481–487. International Joint Conferences on Artificial Intelligence Organization. Main track.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020b. Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4686, Online. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Sentence-level multilingual multi-modal embedding for natural language processing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 139–148, Varna, Bulgaria. INCOMA Ltd.

Jose Camacho-Collados and Roberto Navigli. 2017. BabelDomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228, Valencia, Spain. Association for Computational Linguistics.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*.

Slawomir Dadas. 2023. Opi at semeval-2023 task 1: Image-text embeddings and multimodal information retrieval for visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 155–162, Toronto, Canada. Association for Computational Linguistics.

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.

Sebastian Diem, Chan Jong Im, and Thomas Mandl. 2023. University of hildesheim at semeval-2023 task 1: Combining pre-trained multimodal and generative models for image disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 130–135, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845.

Omid Ghahroodi, Seyed Arshan Dalili, Sahel Mesforoush, and Ehsaneddin Asgari. 2023. Team SUT at SemEval-2023 Task 1: Prompt Generation for Visual Word Sense Disambiguation. In *Proceedings of the*

17th International Workshop on Semantic Evaluation (SemEval-2023), Toronto, Canada. Association for Computational Linguistics.

Małgorzata Grębowiec. 2023. Opi pib at semeval-2023 task 1: A clip-based solution paired with an additional word context extension. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 482–487, Toronto, Canada. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Nikita Katyal, Pawan Rajpoot, Subhanandh Tamilarasu, and Joy Mustafi. 2023. teampn at semeval-2023 task 1: Visual word sense disambiguation using zero-shot multimodal approach. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 457–461, Toronto, Canada. Association for Computational Linguistics.

Jie Li, Yow-Ting Shiue, Yong-Siang Shih, and Jonas Geiping. 2023a. Augmenters at semeval-2023 task 1: Enhancing clip in handling compositionality and ambiguity for zero-shot visual wsd through prompt augmentation and text-to-image diffusion. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 44–49, Toronto, Canada. Association for Computational Linguistics.

Keyi Li, Sen Yang, Chenyang Gao, and Ivan Marsic. 2023b. Rutgers multimedia image processing lab at semeval-2023 task-1: Text-augmentation-based approach for visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1483–1490, Toronto, Canada. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Zhenghui Li, Qi Zhang, XueYin Xia, Yinxiang Ye, Qi Zhang, and Cong Huang. 2023c. Ecnu_miv at semeval-2023 task 1: Ctim - contrastive text-image model for multilingual visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 101–107, Toronto, Canada. Association for Computational Linguistics.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Thanet Markchom, Huizhi Liang, Joyce Gitau, Zehao Liu, Varun Ojha, Lee Taylor, Jake Bonnici, and Abdullah Alshadadi. 2023. UoR-NCL at SemEval-2023 Task 1: Learning Word-Sense and Image Embeddings for Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Antonina Mijatovic, Davide Buscaldi, and Ekaterina Borisova. 2023. RCLN at SemEval-2023 Task 1: Leveraging Stable Diffusion and Image Captions for Visual WSD. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.

Mohammadreza Molavi and Hossein Zeinali. 2023. Slt at semeval-2023 task 1: Enhancing visual word sense disambiguation through image text retrieval using blip. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1921–1925, Toronto, Canada. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Michael Ogezi, Bradley Hauer, Talgat Omarov, Ning Shi, and Grzegorz Kondrak. 2023. UAlberta at SemEval-2023 Task 1: Context Augmentation and Translation for Multilingual Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Rahul Patil, Pinal Patel, Charin Patel, and Mangal Verma. 2023. Rahul patil at semeval-2023 task 1: V-wsd: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1271–1275, Toronto, Canada. Association for Computational Linguistics.

Mohammad Javad Pirhadi, Motahhare Mirzaei, Mohammad Reza Mohammadi, and Sauleh Eetemadi. 2023. PMCoders at SemEval-2023 Task 1: RAltCLIP: Use Relative AltCLIP Features to Rank. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Clifton Poth, Martin Hentschel, Tobias Werner, Hannah Sterz, and Leonard Bongard. 2023. Ml mob at semeval-2023 task 1: Probing clip on visual word-sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1463–1469, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. 2022. DALLE-2 is seeing double: Flaws in word-to-concept mapping in Text2Image models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 335–345, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Florian Schneider and Chris Biemann. 2023. Lt at semeval-2023 task 1: Effective zero-shot visual word sense disambiguation approaches using external knowledge sources. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 462–468, Toronto, Canada. Association for Computational Linguistics.

Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Zeinab Sadat Taghavi, Parsa Haghighi Naeini, Mohammad Ali Sadraei Javaheri, Soroush Gooran, Ehsaneddin Asgari, Hamid Reza Rabiee, and Hossein Sameti. 2023. Ebhaam at semeval-2023 task 1: A clip-based approach for comparing cross-modality and unimodality in visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1960–1964, Toronto, Canada. Association for Computational Linguistics.

Lorenzo Vaiani, Luca Cagliero, and Paolo Garza. 2023. Polito at semeval-2023 task 1: Clip-based visual-word sense disambiguation based on back-translation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1447–1453, Toronto, Canada. Association for Computational Linguistics.

Yuchen Wei and Milton King. 2023. StFX NLP at SemEval-2023 Task 1: Multimodal Encoding-based Methods for Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qihao Yang, Yong Li, Xuelin Wang, Shunhao Li, and Tianyong Hao. 2023. TAM of SCNU at SemEval-2023 Task 1: FCLL: A Fine-grained Contrastive Language-Image Learning Model for Cross-language Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Shibingfeng Zhang, Shantanu Nath, and Davide Mazzaccara. 2023a. GPL at SemEval-2023 Task 1: WordNet and CLIP to Disambiguate Images. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Xudong Zhang, Tiange Zhen, Jing Zhang, Yujin Wang, and Song Liu. 2023b. SRCB at SemEval-2023 Task 1: Prompt Based and Cross-Modal Retrieval Enhanced Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.