UAlberta at SemEval-2023 Task 1: Context Augmentation and Translation for Multilingual Visual Word Sense Disambiguation

Michael Ogezi, Bradley Hauer, Talgat Omarov, Ning Shi, Grzegorz Kondrak
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
`{mikeogezi,bmhauer,omarov,ning.shi,gkondrak}@ualberta.ca`

Abstract

We describe the systems of the University of Alberta team for the SemEval-2023 Visual Word Sense Disambiguation (V-WSD) Task. We present a novel algorithm that leverages glosses retrieved from BabelNet, in combination with text and image encoders. Furthermore, we compare language-specific encoders against the application of English encoders to translated texts. As the contexts given in the task datasets are extremely short, we also experiment with augmenting these contexts with descriptions generated by a language model. This yields substantial improvements in accuracy. We describe and evaluate additional V-WSD methods which use image generation and text-conditioned image segmentation. Overall, the results of our official submission rank us 18 out of 56 teams. Some of our unofficial results are even better than the official ones. Our code is publicly available at https://github.com/UAlberta-NLP/v-wsd.

1 Introduction

This paper addresses our work on SemEval-2023 Task 1: Visual Word Sense Disambiguation[1] (Raganato et al., 2023). The V-WSD task is closely related to WSD, and similarly involves understanding and classifying the meaning of a polysemous word in context. The distinction is in how classes are defined: In WSD, a system has access to a sense inventory that enumerates the possible senses of each word, and the task is to classify the focus word according to the sense that best corresponds to its intended meaning. In V-WSD, a system is given a set of candidate images, and the task is to select the image which depicts the intended meaning of the focus word.

The multi-modal nature of V-WSD introduces challenges not encountered in WSD. First, image

---

[1] https://raganato.github.io/vwsd/

Figure 1: The task is to select the image that best represents the meaning of the focus word (e.g., *bat*) in the context (e.g., "baseball bat.")

processing is generally more computationally intensive than text processing. Second, a V-WSD system must represent the meanings of both images and text, and must have mechanisms to compare these multi-modal semantic representations. Last, since the candidate images in V-WSD are not restricted to a sense inventory, they may exhibit highly variable levels of sense granularity.

The V-WSD task is motivated by cases where textual context alone is insufficient to disambiguate a word. In such cases, visual context may be available to facilitate disambiguation. For example, the word *play* is ambiguous in the context "that was a good play," as it may refer to a theatrical performance or an action in a sport. However, an associated image of a stage or a sports field would enable a V-WSD system to disambiguate *play*.

We propose a novel V-WSD algorithm that ranks candidate images by embedding images and words-in-context in a shared semantic space, while also taking advantage of lexical knowledge bases commonly used in WSD. In particular, our method uses sense glosses of the focus word to create representations of the possible meanings that word may have. Our algorithm is flexible, It includes several optional modules, as well as hyper-parameters that facilitate customization, optimization, and detailed analysis.

We test various configurations of our method and analyze their performance. Our three principal conclusions are as follows: First and foremost, the

augmention of the original textual context plays a crucial role in improving performance. Second, there is a considerable gap between English and non-English performance, indicating that bias towards English models extends to the multi-modal setting. Third, we observe a major distribution shift between the train and test sets, which is confirmed by our ablation study.

2 Related Work

Recent work on WSD can be divided into supervised and knowledge-based systems. Supervised WSD methods depend on large training corpora in which some or all of the content words have been tagged with their correct senses (Blevins and Zettlemoyer, 2020; Barba et al., 2021). Knowledge-based methods depend on other sources of linguistic knowledge (Wang and Wang, 2020). In general, knowledge-based methods are outperformed by contemporary supervised methods (Pasini et al., 2021). Today, state-of-the-art WSD systems approach accuracy limits imposed by inter-annotator agreement (Maru et al., 2022).

Research on the incorporation of visual information for WSD is relatively sparse. Barnard et al. (2003) propose a statistical model that associates image regions and words to predict word senses. Loeff et al. (2006) apply spectral clustering to group similar images corresponding to the same senses. Saenko and Darrell (2008) employ an unsupervised approach to assign senses to images using surrounding texts and dictionary definitions, and then train a visual SVM classifier to disambiguate unseen images. Gella et al. (2019) introduce the task of visual verb sense disambiguation, in which one image is selected based on a given context. Vascon et al. (2021) propose a graph-based semi-supervised transductive learning method for visual verb sense disambiguation.

Multi-modal foundation models (Bommasani et al., 2021) such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) can represent both text and images in a shared embedding space. Recent work[2] (Bianchi et al., 2021; Sajjad Ayoubi, 2022) improves the text encoder by using a pretrained text-only encoder such as BERT (Devlin et al., 2018).

---

3 Task & Dataset

Task Definition: Given a focus word w in a short context c, and a set of candidate images I, the task is to select the image $i^* \in I$ which best represents the meaning of w in c. For example, given the context "baseball bat" with *bat* as the focus word, a V-WSD system should choose the image that depicts the bat used in baseball (Figure 1).

Dataset: The training data provided for this shared task consists of a silver dataset with 12,869 V-WSD instances. Each sample is a 4-tuple $\langle f, c, I, i^* \in I \rangle$ where $|I| = 10$. The contexts are generally very short, often just a single word in addition to the focus word. We randomly select 10% of the training data for use as a development set. The test dataset consists of 968 instances, of which 463 are English, 200 are Farsi, and 305 are Italian. We observe that many of the incorrect candidate images in the training data have nothing to do with any sense of the focus word. However, in the test data, we observe that this is less often the case, making the test set considerably more difficult.

Evaluation Metrics: The primary metric is the hit rate, which is equivalent to top-1 accuracy, or simply accuracy. This is the proportion of instances for which the system selects the correct image. We also compute the mean reciprocal rank (Voorhees and Tice, 2000) which represents how highly V-WSD systems rank the ground-truth image, on average.

Language	%	Example
English	82	waxflower wildflower
Latin	15	shorea genus
German	2	truppenübungsplatz workplace
French	1	brumaire month

Table 1: The language distribution of 100 instances from the training set. The focus word is underlined.

Language Distribution: We observed some instances where the context contained non-English words. To estimate the prevalence of this phenomenon, we randomly selected 100 instances from the training set and manually identified the language of each. For example, the focus word *shorea* in "shorea genus" comes from new Latin, and refers to a genus of mainly rainforest trees. Table 1 shows the frequency of each language in our sample.

4 Method

In this section, we describe the key components of our systems, including an algorithm that combines text and image similarity measures.

4.1 Algorithm

We propose an algorithm to select a single image from a set of candidates that best matches the context. To reiterate the problem, we are given a context c containing a focus word w and a set I of candidate images. We assume that we also have a non-empty set G containing possible glosses of w; in practice, we obtain G from BabelNet using the freely available API.[3]

Our algorithm makes calls to two similarity functions: The first is sim^L, a *written language* similarity function, which takes as input two text strings and returns a value indicating the semantic similarity between them. The second is sim^{VL}, a *vision-to-written language* similarity function, which takes as input an image and a text string and returns a value indicating the similarity between what the image depicts and what the text describes.

With these functions, for each candidate image $i \in I$, and for each gloss $g \in G$ of the focus word w, we compute the pairwise similarity between:

1. The image and context: $s_{ic} = sim^{VL}(i, c)$

2. The image and gloss: $s_{ig} = sim^{VL}(i, g)$

3. The context and gloss: $s_{cg} = sim^L(c, g)$

This allows us to identify the pair of a candidate image i^* and gloss g^* that maximizes a weighted average of these three similarity scores. Algorithm 1 shows the pseudocode for this algorithm.

Hyperparameters: Our algorithm depends on three weight hyperparameters: w_{ic}, w_{ig}, and w_{cg}. They represent the weights for image-context, image-gloss, and context-gloss similarity, respectively. Table 2 shows the results of the hyperparameter binarized grid search performed on a 500-sample of the training set. Based on our development experiment results, we decided to set all hyperparameter weights to 1 for simplicity, except where otherwise noted. We discuss the hyperparameters further in Section 6.6.

---

[3]https://babelnet.org/guide

Algorithm 1 Candidate Image Scoring

1: $c \leftarrow$ the context of the focus word
2: $G \leftarrow$ list of glosses for the focus word
3: $I \leftarrow$ list of candidate images
4: **for** i in I **do**
5:      $s_g \leftarrow 0$
6:      **for** g in G **do**
7:          $s_{ig} \leftarrow w_{ig} \cdot sim^{VL}(i, g)$
8:          $s_{cg} \leftarrow w_{cg} \cdot sim^L(c, g)$
9:          $s_g \leftarrow \max(s_g, s_{ig} + s_{cg})$
10:      $scores[i] \leftarrow s_g + w_{ic} \cdot sim^{VL}(i, c)$
11: **return** $scores$

Context Augmentation: For each instance, we prompt InstructGPT (Brown et al., 2020; Ouyang et al., 2022) to generate a definition for the context phrase. We use the following prompt template: "For each line, define the phrase:" followed by the contexts, one per line. For example, the context "baseball bat" is augmented to become "baseball bat: a bat used to hit a baseball during the game of baseball." The use of this additional context is described in Section 5.3

Supplementary Training Data: We speculate that the size of the training dataset may be a limiting factor in the accuracy of our method. We, therefore, experiment with augmenting the training data with additional data derived from BabelPic Calabrese et al. (2020), a multi-modal resource which maps a subset of BabelNet synsets to sets of one or more images. For each pair of a synset and an image, we enumerate a lemma from the base synset and a lemma from a related synset. The two lemmas are concatenated, starting with the lemma from the related base synset, to form a two-word context. We then select nine other random images from BabelPic, forming an instance comparable to those in the training set: a two-word context with a

w_{ic}	w_{ig}	w_{cg}	Accuracy (%)
1	1	1	79.2
1	1	0	79.2
1	0	1	72.2
1	0	0	72.2
0	1	1	68.4
0	1	0	68.6
0	0	1	11.0

Table 2: Binarized grid search results for weight hyperparameters.

single focus word, with ten images, one depicting the correct sense of the focus. We create 54,968 instances this way and experiment with adding this dataset to the training data at training time.

Glosses: For each instance, we enumerate the BabelNet (Navigli and Ponzetto, 2012) glosses corresponding to each sense of the focus word. If there are multiple glosses for a single sense, we pick the first and add it to the set G. This prevents senses from being over-represented due to the number of glosses in BabelNet.

5 Systems

In this section, we describe our systems for the V-WSD task, Our official system submissions are based on our primary systems: TR and LANGSPEC. We also describe two alternative systems, which do not use Algorithm 1. Both perform worse than the primary systems, but their results are nevertheless valuable for the purpose of analysis. We also present a supplementary method, which can be optionally used in combination with our other systems. Non-English instances are translated using DeepL[4] for Italian and ChatGPT[5] for Farsi.

5.1 Primary Systems

TR: Image Scoring with Translations If the input instance is not English, we translate it into English. Then we apply Algorithm 1. We compute sim^{VL} using embeddings from CLIP (Radford et al., 2021), an English-only model which encodes text and images in a shared embedding space. We compute sim^L using BERT (Devlin et al., 2018) as an English-only text encoder. We set the weight parameters: w_{ic}, w_{ig}, and w_{cg} to 1 in this specific case.

LANGSPEC: Image Scoring with Language-Specific Models This system is similar to TR, except that non-English instances are not translated into English. This is our only system which directly operates in other languages. Given a non-English instance, we replace CLIP and BERT with language-specific models to compute sim^{VL} and sim^L. For English instances, this method is the same as TR. For Italian, we use CLIP-Italian (Bianchi et al., 2021) to compute sim^{VL} and Italian BERT[6] to compute sim^L. For Farsi, we use

CLIPfa (Sajjad Ayoubi, 2022) to compute sim^{VL} and ParsBERT (Farahani et al., 2021) to compute sim^L.

5.2 Alternative Systems

GEN: Generative Image Model This method takes a different approach compared to TR and LANGSPEC; it does not use Algorithm 1. Instead, we provide the context (translated into English, if needed, as outlined above) as input to Stable Diffusion (Rombach et al., 2022), a generative model which takes a text prompt as input and produces candidate images to depict what the text describes. For each context, we generate 15 images using 20 diffusion steps each. We set the guidance scale hyperparameter to 7.5. For each candidate image, we compute its cosine similarity with each generated image based on embeddings produced by CLIP. The candidate with the highest similarity to the generated images is chosen as the output.

SEG: Text-Conditioned Image Segmentation As with GEN, this method does not use Algorithm 1. Instead, we use a zero-shot image segmentation system (Lüddecke and Ecker, 2022) to segment images based on the provided context. This system produces a *mean mask value*, which we use as a measure of similarity between the context and the segmented image; we return the image with the highest mean mask value, given the context.

5.3 Supplementary Method

DEF: Generating Additional Context TR, GEN, and SEG make use of the input context, translated to English as needed. However, the contexts provided in the official dataset for this task are extremely short. With DEF, we generate additional context by using the original context to prompt InstructGPT for a more extensive description, as described in Section 4.1. We then concatenate the generated text to the context and pass this augmented context to TR, GEN, or SEG. We refer to the methods using this supplementary method as TR+DEF, GEN+DEF, and SEG+DEF, respectively. We do not combine DEF with LANGSPEC, as we observe that InstructGPT is less robust to short non-English contexts.

For TR+DEF, we set w_{ig} and w_{cg} to 0, as the improved context obviates the need for their corresponding terms in Algorithm 1. GEN+DEF and SEG+DEF, being based on GEN and SEG, do not depend on Algorithm 1.

---

[4] https://www.deepl.com/translator
[5] https://openai.com/blog/chatgpt/
[6] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

	EN	IT	FA	Avg
Baseline	60.5	22.6	28.5	37.2
TR*	61.1	59.3	**43.0**	54.5
TR+DEF	**69.1**	**63.3**	40.0	**57.5**
LANGSPEC*	56.8	37.7	14.5	36.3
GEN	51.6	45.9	39.0	45.5
GEN+DEF	58.1	48.5	34.5	47.0
SEG	31.5	29.8	20.5	27.3
SEG+DEF	34.1	36.7	20.0	30.3

Table 3: Accuracy for English, Italian, and Farsi, along with the macro average for all languages. We indicate our official system submissions with *.

6 Experiments

In this section, we present, discuss, and analyze our results.

6.1 Results

Table 3 shows our performance on the test set. We find that accuracy has a 99.46% Pearson correlation with mean reciprocal rank, and so for conciseness, we report accuracy alone. The translation-based systems, TR and TR+DEF, yield the best results. One explanation for this outcome is the disproportionate amount of English training data available to the models we build upon: CLIP and BERT. The higher performance of these models on English appears to compensate for the noise introduced by the translation process. We discuss this further in Section 6.4.

An interesting trend is the benefit of context augmentation, (Section 5.3). Between TR and TR+DEF, we observe a 3% average improvement in accuracy. We observe a similar trend in GEN versus GEN+DEF and SEG versus SEG+DEF.

We further observe that accuracy on English instances is highest, accuracy on Farsi instances is lowest, while accuracy on Italian instances is in between both. This corresponds to the quality and quantity of resources available for each language. We undertake more thorough analyses in the next section.

6.2 Distribution Shift

As shown in Table 4, we observe a clear disparity in polysemy, and the proportion of focus words which are nouns, between the training and test sets. This difference is especially notable when considering the performance gap between the sets.

	Train	Test		
	EN	EN	IT	FA
Polysemy	6.8	23.1	13.6	10.7
Nouns (%)	74.7	88.1	91.5	92.5

Table 4: Distribution shifts between the training and test sets. Polysemy indicates the average number of senses each focus word in the set has.

Zero-shot vs. Fine-tuning: We observe that fine-tuning on the training set leads to a drop in performance on the test set (Figure 2). This may be due to the divergence between the training and test datasets outlined above.
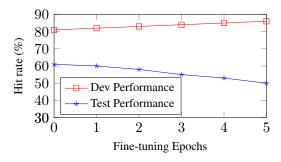

Figure 2: As we fine-tune on the training set for more epochs, we see an increase in dev set performance, but a drop in test set performance. Epoch 0 refers to using the model zero-shot, with no fine-tuning.

6.3 Traditional WSD

Although both the V-WSD and WSD tasks have some similarities, we found that some ideas drawn from WSD prove ineffective for V-WSD.

Using Glosses: We observe empirically worse performance when using glosses in our algorithm. Specifically, with TR on the English test set, we obtain a hit rate of 61.1% when we do not use glosses and 56.8% when we do. Such a steep drop (4.3%) is surprising, especially since most state-of-the-art WSD systems explicitly use glosses in their methods.

We posit that sense disambiguation in V-WSD is more focused on homonymy than polysemy and, as a result, can be less nuanced than in WSD. For example, *apple* could refer to a fruit or a tree. In an image depicting both, the focus may be unclear. In WSD, this distinction is critical since *tree* and *fruit* are distinct senses. In V-WSD, however, we can make a correct prediction without deciding between both senses. As a result of this lower granularity, glosses become less important.

Performance of WSD Systems on Context: We manually disambiguated the sense of the focus word in a randomly-selected set of 16 instances from the training set. We then applied a state-of-the-art WSD system, ConSec (Barba et al., 2021), to these instances. We observe that ConSec sense predictions were accurate 50% of the time, falling considerably below its reported accuracy of 82%.

6.4 English Hegemony

Natural language processing research often focuses on the English language, at the expense of other languages (Magueresse et al., 2020). The relative performance of TR and LANGSPEC reflects this phenomenon: Translating non-English text to English, in order to apply an English encoder, can be expected to introduce some noise due to translation errors and information loss. However, we observe that this pipeline approach produces better results than using an Italian or Farsi encoder directly. This suggests that the field's focus on English has yielded English encoders which are much better than those available for Italian or Farsi. We speculate that advancing the state-of-the-art for non-English encoders may yield even better performance, by avoiding the need to translate to English.

6.5 Image Generation

As shown in Figure 3, when applying our image generation system (GEN), we observe an increase in performance as we generate more images. Although the performance jump when transitioning from 1 to 5 images is most pronounced, we see benefits from scaling until a certain point, 10 images, where the trend becomes unreliable.

Figure 3: Hit rate (%) vs. number of images generated for GEN and GEN+DEF.

6.6 Text-Conditioned Image Segmentation

With SEG, we can sometimes robustly segment images and predict masks indicating the correct
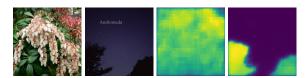

Figure 4: Original images from the dataset depicting ANDROMEDA (Japanese plant), ANDROMEDA (galaxy) and their two masks conditioned on "andromeda tree."

image, conditioning on the full context. However, this method sometimes forms incorrect semantic representations. Appendix A details more examples of SEG's usage. In addition to Figure 4, we present more extensive examples in Appendix A.

6.7 Algorithm Hyperparameters

Our algorithm uses three weights hyperparameters to balance pairwise similarities. We set all weights to 1 based on Table 2. Comparison of results with w_{cg} set to 0 or 1 suggests that $sim^L(c, g)$ does not improve performance. Two reasons support this finding. Firstly, images encode richer representations, producing more precise $sim^{VL}(i, g)$ and $sim^{VL}(i, c)$, while both context and glosses are discrete textual features, introducing uncertainty to $sim^L(c, g)$. Secondly, we use CLIP and BERT to calculate sim^{VL} and sim^L, respectively. CLIP's multi-modal pre-training may offer better similarity scores, fitting this task better. Understanding these findings more deeply is an interesting avenue for future research.

7 Conclusion

In this paper, we outlined our work on the recently proposed task of V-WSD. We found that many ideas from traditional WSD are difficult to adapt to V-WSD, and, moreover, WSD systems are generally not useful for V-WSD. We were particularly surprised to find that, unlike in WSD, glosses appear to be unhelpful for V-WSD. Contrariwise, our innovation of augmenting the context did yield substantial gains in accuracy.

Further research will be needed to establish the connection between V-WSD and the broader field of lexical semantics. We speculate that developing systems for joint WSD and V-WSD may yield improvements in one or both tasks. Our work here serves as a proof-of-concept establishing the utility of language models and lexico-semantic resources in the developing task of V-WSD.

Acknowledgements

References

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kobus Barnard, Matthew Johnson, and David Forsyth. 2003. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data*, pages 1–5.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pretraining for the italian language. *arXiv preprint arXiv:2108.08688*.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*.

Spandana Gella, Frank Keller, and Mirella Lapata. 2019. Disambiguating visual verbs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):311–322.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 547–554, Sydney, Australia. Association for Computational Linguistics.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of Word Sense Disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. 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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and crosslingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

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.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 Task 1: Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Kate Saenko and Trevor Darrell. 2008. Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Amir Ahmadi Sajjad Ayoubi, Navid Kanaani. 2022. Clipfa: Connecting farsi text and images. https://github.com/SajjjadAyobi/CLIPfa.

Sebastiano Vascon, Sinem Aslan, Gianluca Bigaglia, Lorenzo Giudice, and Marcello Pelillo. 2021. Transductive visual verb sense disambiguation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3049–3058.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.

A Text-Conditioned Image Segmentation

A.1 Success Mode

In the successful case of this system, we see that we are able to segment the object based on the text provided properly. See the figures below for details.
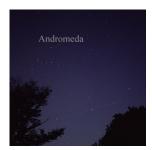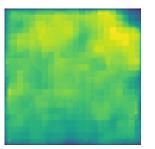

Figure 5: Original images from the dataset depicting on the left: ANDROMEDA and on the right: ANDROMEDA.
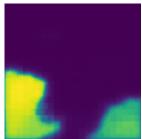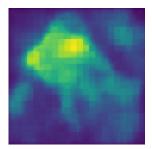

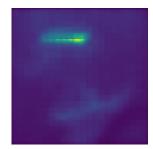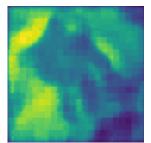Figure 6: Conditioned on the full "andromeda tree"
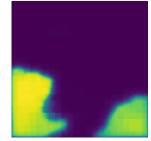

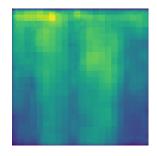Figure 7: Conditioned on "andromeda"

Figure 8: Conditioned on "tree"

A.2 Failure Mode

In the failure case of this system, we see that we are unable to confidently segment the object based on the text provided. See the figures below for details.

Figure 9: Original images from the dataset depicting on the left: BANK (finance) and on the right: BANK (river).
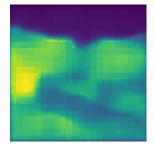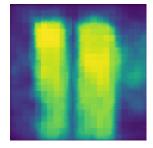

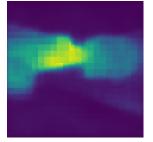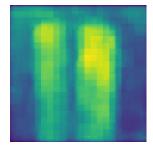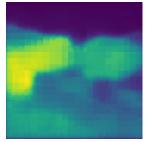Figure 10: Conditioned on the full "bank erosion"

Figure 11: Conditioned on "bank"

Figure 12: Conditioned on "erosion"