PolCLIP: A Unified Image-Text Word Sense Disambiguation Model via Generating Multimodal Complementary Representations

Qihao Yang[1], **Yong Li**[1], **Xuelin Wang**[2], **Fu Lee Wang**[3], **Tianyong Hao**[1]*
[1]School of Computer Science, South China Normal University, Guangzhou, China
[2]College of Chinese Language and Culture, Jinan University, Guangzhou, China
[3]School of Science and Technology, Hong Kong Metropolitan University, Hong Kong
charlesyeung@m.scnu.edu.cn

Abstract

Word sense disambiguation (WSD) can be viewed as two subtasks: textual word sense disambiguation (Textual-WSD) and visual word sense disambiguation (Visual-WSD). They aim to identify the most semantically relevant senses or images to a given context containing ambiguous target words. However, existing WSD models seldom address these two subtasks jointly due to lack of images in Textual-WSD datasets or lack of senses in Visual-WSD datasets. To bridge this gap, we propose PolCLIP, a unified image-text WSD model. By employing an image-text complementarity strategy, it not only simulates stable diffusion models to generate implicit visual representations for word senses but also simulates image captioning models to provide implicit textual representations for images. Additionally, a disambiguation-oriented image-sense dataset is constructed for the training objective of learning multimodal polysemy representations. To the best of our knowledge, PolCLIP is the first model that can cope with both Textual-WSD and Visual-WSD. Extensive experimental results on benchmarks demonstrate the effectiveness of our method, achieving a 2.53% F1-score increase over the state-of-the-art models on Textual-WSD and a 2.22% HR@1 improvement on Visual-WSD.

1 Introduction

Understanding and identifying the intended meaning of words with multiple senses (i.e., polysemy) is a significant challenge in natural language processing (Navigli, 2009). This promotes in-depth research on word sense disambiguation (WSD), which has recently been extended to multimodal downstream tasks (Bevilacqua et al., 2021). Techniques for WSD are critical for enhancing the accuracy and effectiveness of text understanding and information retrieval tasks such as machine translation (Raganato et al., 2019), image-text retrieval (Chen et al., 2020), and large language model inference (Kritharoula et al., 2023).

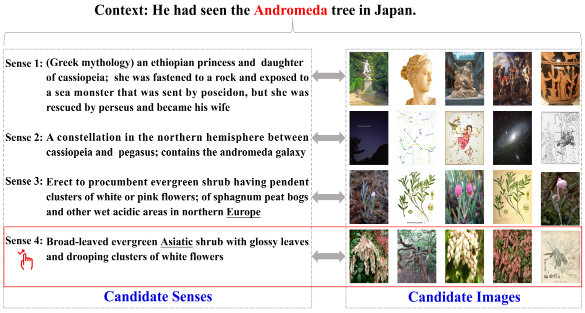

Context: He had seen the Andromeda tree in Japan.
Sense 1: (Greek mythology) an ethiopian princess and daughter of cassiopeia; she was fastened to a rock and exposed to a sea monster that was sent by poseidon, but she was rescued by perseus and became his wife
Sense 2: A constellation in the northern hemisphere between cassiopeia and pegasus; contains the andromeda galaxy
Sense 3: Erect to procumbent evergreen shrub having pendent clusters of white or pink flowers; of sphagnum peat bogs and other wet acidic areas in northern Europe
Sense 4: Broad-leaved evergreen Asiatic shrub with glossy leaves and drooping clusters of white flowers
Candidate Senses
Candidate Images

Figure 1: Illustration of the Multimodal-WSD task.

Theoretically, WSD can be divided into two subtasks: textual word sense disambiguation (Textual-WSD) (Bevilacqua et al., 2021) and visual word sense disambiguation (Visual-WSD) (Raganato et al., 2023). Given a context containing an ambiguous target word, the goal of Textual-WSD is to select the most semantically appropriate one from a set of candidate senses, while the goal of Visual-WSD is to choose the most semantically suitable one from a set of candidate images. Due to the distinct modalities, these two subtasks typically require specialized training datasets and methods (Bevilacqua and Navigli, 2020; Blevins and Zettlemoyer, 2020; Kwon et al., 2023). Nevertheless, they can be unified as a Multimodal-WSD task if the senses and images in existing WSD datasets are aligned, as shown in Figure 1. The task objectives of Multimodal-WSD cover: 1) in cases where candidates include both senses and images, a Multimodal-WSD model is required to identify the most semantically relevant senses and images simultaneously for ambiguous target words within a given context; 2) in cases where candidates include either senses or images, the model is still

*Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10676–10690
August 11-16, 2024 ©2024 Association for Computational Linguistics

required to identify the most semantically relevant senses or images for ambiguous target words within a given context. Such task objectives are also in line with the practical scenarios of current large vision-language models (Liu et al., 2024). Technically, developing a generic Multimodal-WSD model can realize the unification of WSD tasks and activate the potential of multimodal applications in understanding polysemy knowledge.

In the Textual-WSD datasets (Raganato et al., 2017), merely textual senses serve as candidates (i.e., image-missing), while in the Visual-WSD datasets (Raganato et al., 2023), only images serve as candidates (i.e., sense-missing). This results in the challenge of modality missing at the data level, limiting the unification of these two WSD subtasks. Furthermore, multimodal representations have been demonstrated to carry richer semantic information compared to unimodal representations in recent WSD works (Gella et al., 2016, 2019). However, constrained by model architecture, existing Textual-WSD models (Conia and Navigli, 2021; Maru et al., 2019; Huang et al., 2019) cannot supplement candidate senses with image information, and Visual-WSD models (Yang et al., 2023; Zhang et al., 2023; Dadas, 2023) cannot supplement candidate images with descriptions. This poses technical difficulties in developing a unified framework for Multimodal-WSD.

To address these issues, we propose **PolCLIP**, a unified image-text WSD model which is proficient in multimodal **pol**ysemy processing and is built upon **CLIP** (Radford et al., 2021) architecture. By employing an image-text complementarity strategy, it simulates stable diffusion models (Ho et al., 2020) (generating images based on texts) and image captioning models (Ramos et al., 2023) (generating descriptions based on images). The core idea of this strategy is to make PolCLIP initially focus on the key information of original unimodal senses or images, and then re-utilize the text or image encoder to generate implicit image-text complementary representations. Two widely used WSD datasets (SemCor (Miller et al., 1993) and VWSD-KB (Yang et al., 2023)) are integrated into a disambiguation-oriented image-sense dataset for the training objective of learning aligned multimodal representations. Moreover, a fine-tuned GPT-3.5 model is utilized to generate lexical definitions for semantic enhancement in testing phase. The main contributions of this work are summarized as follows:

- A unified image-text WSD model is proposed, which is the first model to jointly cope with Textual-WSD lacking images and Visual-WSD lacking senses.
- An image-text complementarity strategy is introduced to simulate stable diffusion models and image captioning models for addressing the modality missing issues in unimodal WSD datasets.
- A disambiguation-oriented image-sense dataset is constructed to provide a benchmark for the Multimodal-WSD task.

2 Related Work

Textual-WSD was mainly tackled by knowledge-based methods and supervised methods (Bevilacqua et al., 2021). Knowledge-based methods (Maru et al., 2019; Scozzafava et al., 2020) typically used external dictionary resources to provide sense lists for ambiguous words to resolve polysemy. Supervised methods (Huang et al., 2019; Wang and Wang, 2020) generally used pre-training language models to maximize the similarity probabilities between contexts and candidate senses in a feature space. BEM (Blevins and Zettlemoyer, 2020) and SACE (Wang and Wang, 2021) adopted bi-encoders and only retained the representations corresponding to ambiguous words. They achieved state-of-the-art results on English all-words benchmarks at that time. Moreover, the full utilization of visual features for verb sense disambiguation has attracted increasing interest (Gella et al., 2016, 2019). EViLBERT (Calabrese et al., 2020b) obtained better results by learning task-agnostic multimodal sense representations, compared to methods built solely on language models. Although these methods primarily leveraged visual information to bolster performance on Textual-WSD, they could not be applied to Visual-WSD straightforwardly.

Visual-WSD was introduced in SemEval-2023 Task 1 (Raganato et al., 2023). The Visual-WSD mainstream approaches employed Vision-Language Pre-training models (VLPs) for image-text retrieval. FCLL (Yang et al., 2023) proposed a fine-grained image-text contrastive learning mechanism and won first place in SemEval-2023 Task 1. Moreover, large language models (LLMs) were widely used to enrich the semantic information of contexts (Ghahroodi et al., 2023; Yang et al., 2024). Calling APIs was a commonly adopted strategy, where simple prompts were designed to

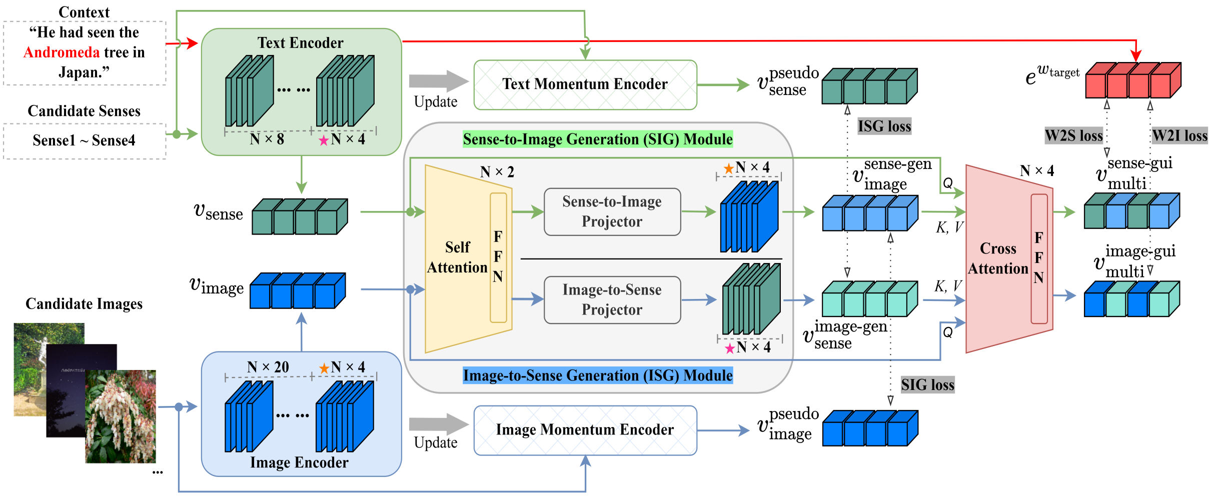


Figure 2: Overview framework of the proposed PolCLIP model.

guide LLMs to return the interpretations of ambiguous target words in contexts (Kritharoula et al., 2023). However, Visual-WSD models depended on the prior knowledge of VLPs, which were pretrained with objectives biased towards image-text understanding rather than WSD. This resulted in Visual-WSD models struggling to effectively resolve Textual-WSD.

Data was a key factor to unify these two WSD subtasks and develop a generic Multimodal-WSD model. WordNet (Miller et al., 1990) was a large lexicographic database and a standard inventory for English WSD. It contained approximately 120,000 synsets. BabelNet (Navigli and Ponzetto, 2010; Navigli et al., 2021) was the most popular multilingual dictionary, which was semi-automatically mapped to other resources to acquire encyclopedic terms. It covered over 500 languages and was upgraded to version 5.3 recently. Researchers could access various possible resources about ambiguous words by BabelNet, including example sentences, parts of speech, textual senses, and images. By linking WordNet with Wikipedia through BabelNet, BabelPic (Calabrese et al., 2020a) expanded non-concrete image-sense pairs, paving the way for our work to construct larger disambiguation-oriented multimodal datasets.

Inspired by FCLL, we used a dual-stream architecture with contrastive learning to align the knowledge of word-sense pairs and word-image pairs from the proposed disambiguation-oriented dataset. Previous research viewed LLMs as common tools for text semantic enhancement. Thus, LLMs were also adopted in our testing phase.

3 Method

3.1 Task formulation

Textual-WSD and Visual-WSD can be unified as a Multimodal-WSD task which is a token classification problem. A given context c generally contains at least one ambiguous target word w_{target}. For Textual-WSD, there is a set of word senses $S = \{s_1, s_2, \ldots, \hat{s}, \ldots, s_n\}$ as candidates, where $\hat{s}$ denotes the most semantically relevant sense to w_{target}. Following Eq. 1, a Textual-WSD model is required to learn a similarity function F to retrieve $\hat{s}$ from candidate senses. For Visual-WSD, there is a group of images $I = \left\{i_1, i_2, \ldots, \hat{i}, \ldots, i_n\right\}$ as candidates, where $\hat{i}$ represents the most semantically relevant image to w_{target}. Following Eq. 2, a Visual-WSD model is required to learn a similarity function F to retrieve $\hat{i}$ from candidate images.

$$\hat{s} = \arg\max F\left(c, w^{target}, S\right) \tag{1}$$

$$\hat{i} = \arg\max F\left(c, w^{target}, I\right) \tag{2}$$

3.2 The PolCLIP model

The framework of the PolCLIP model is shown in Figure 2. It utilizes an image-text complementarity strategy and is built upon CLIP architecture. It employs 12-layer transformers as the text encoder and 24-layer visual transformers as the image encoder. A context with an ambiguous target word w_{target} is input into the text encoder to generate a complete context representation $e_c = \{[CLS], e^{w_1}, \ldots, e^{w_{\text{target}}}, \ldots, e^{w_n}, [SEP]\}$, where $e^{w_{\text{target}}}$ is the representation corresponding to w_{target}. The candidate senses of w_{target} are input into the text encoder to output a sense vector v_{sense}. The candidate images of w_{target} are fed into the image encoder to output an image vector v_{image}.

A Sense-to-Image Generation (SIG) module and an Image-to-Sense Generation (ISG) module are designed to generate implicit image-text complementary information, which is expressed in vectors. Specifically, the SIG module consists of a shared 2-layer self-attention module, a sense-to-image projector, and the last four layers of the image encoder. The sense vector v_{sense} is input to SIG. Its key information is condensed through the self-attention module and the linear-layer projector. After that, the highly compressed sense information is transformed into a sense-generated image vector $v_{\text{image}}^{\text{sense-gen}}$ with visual knowledge through the last four layers of the image encoder. Similarly, the ISG module is composed of the same shared 2-layer self-attention module, an image-to-sense projector, and the last four layers of the text encoder. The image vector v_{image} is input to ISG and then converted into an image-generated sense vector $v_{\text{sense}}^{\text{image-gen}}$ with textual knowledge.

In order to make the generated implicit image-text complementary information actually beneficial to semantic augment, following ALBEF (Li et al., 2021), a text momentum encoder and an image momentum encoder are employed to generate pseudo-target vectors. Specifically, the candidate senses and images from the same batch are input separately into the text and image momentum encoders to output a pseudo-target sense vector $v_{\text{sense}}^{\text{pseudo}}$ and a pseudo-target image vector $v_{\text{image}}^{\text{pseudo}}$. These two pseudo-target vectors supervise $v_{\text{sense}}^{\text{image-gen}}$ and $v_{\text{image}}^{\text{sense-gen}}$ to be close to ground truth. The text and image momentum encoders retain the prior knowledge of the backbone model to counteract the issue of catastrophic forgetting (Li et al., 2023). Therefore, the pseudo-target vectors gain improvement continuously with these two momentum encoders being optimized at a small pace. The similarity between $v_{\text{sense}}^{\text{image-gen}}$ and $v_{\text{sense}}^{\text{pseudo}}$ is calculated by Eq. 3-4. Also, the similarity between $v_{\text{image}}^{\text{sense-gen}}$ and $v_{\text{image}}^{\text{pseudo}}$ is calculated by Eq. 5-6. s is a similarity function. $\mathcal{P}^{ISG}$ and $\mathcal{P}^{SIG}$ are the softmax-normalized similarities used to supervise the ISG module and the SIG module.

$$s\left(S^{\text{gen}}, S^{\text{pse}}\right) = v_{\text{sense}}^{\text{image-gen}} \cdot \left(v_{\text{sense}}^{\text{pseudo}}\right)^{\text{T}} \tag{3}$$

$$\mathcal{P}^{ISG} = \frac{\exp\left(s\left(S^{\text{gen}}, S^{\text{pse}}\right)\right)}{\sum_{\text{n}=1}^{N} \exp\left(s\left(S^{\text{gen}}, S^{\text{pse}}\right)\right)} \tag{4}$$

$$s\left(I^{\text{gen}}, I^{\text{pse}}\right) = v_{\text{image}}^{\text{sense-gen}} \cdot \left(v_{\text{image}}^{\text{pseudo}}\right)^{\text{T}} \tag{5}$$

$$\mathcal{P}^{SIG} = \frac{\exp\left(s\left(I^{\text{gen}}, I^{\text{pse}}\right)\right)}{\sum_{\text{n}=1}^{N} \exp\left(s\left(I^{\text{gen}}, I^{\text{pse}}\right)\right)} \tag{6}$$

A shared 4-layer cross-attention module serves as a fusion module. It integrates the original unimodal sense/image representations and the generated implicit image/sense representations into semantically enriched multimodal representations. v_{sense} serves as Q and $v_{\text{image}}^{\text{sense-gen}}$ serves as K and V. They are fed into the fusion module and then a sense-guided multimodal vector $v_{\text{multi}}^{\text{sense-gui}}$ is calculated by $\text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}V\right)$, where d_k denotes the dimension of 768. This $v_{\text{multi}}^{\text{sense-gui}}$ achieves an effective interaction of the original sense representation with the implicit sense-generated image representation. Similarly, v_{image} serves as Q and $v_{\text{sense}}^{\text{image-gen}}$ serves as K and V. They are fed into the fusion module and then an image-guided multimodal vector $v_{\text{multi}}^{\text{image-gui}}$ is output by the same cross-attention calculation process. This $v_{\text{multi}}^{\text{image-gui}}$ achieves an effective interaction of the original image representation with the implicit image-generated sense representation.

To avoid the key information of the ambiguous target word w_{target} being smoothed out, we directly select the representation of w_{target} as the anchor vector for retrieval, instead of simply averaging the complete context representation e_c or taking $[CLS]$. Following Eq. 7-8, this anchor $e^{w_{\text{target}}}$ is used to calculate the similarity with the sense-guided multimodal vector $v_{\text{multi}}^{\text{sense-gui}}$. The similarity between the anchor $e^{w_{\text{target}}}$ and the image-guided multimodal vector $v_{\text{multi}}^{\text{image-gui}}$ is calculated by Eq. 9-10. $\mathcal{P}^{W2S}$ and $\mathcal{P}^{W2I}$ are the softmax-normalized anchor-to-sense similarity and the anchor-to-image similarity. The PolCLIP model can identify the most semantically appropriate senses and images based on these two similarities.

$$s(W, S) = e^{w_{\text{target}}} \cdot \left(v_{\text{multi}}^{\text{sense-gui}}\right)^{\text{T}} \tag{7}$$

$$\mathcal{P}^{W2S} = \frac{\exp(s(W, S))}{\sum_{\text{n}=1}^{N} \exp(s(W, S))} \tag{8}$$

$$s(W, I) = e^{w_{\text{target}}} \cdot \left(v_{\text{multi}}^{\text{image-gui}}\right)^{\text{T}} \tag{9}$$

$$\mathcal{P}^{W2I} = \frac{\exp(s(W, I))}{\sum_{\text{n}=1}^{N} \exp(s(W, I))} \tag{10}$$

Algorithm 1: Pseudocode of Training PolCLIP

```
data : a context $c$ with an ambiguous target word $w_{\text{target}}$;
       the candidate senses $S$ and the candidate images $I$;
while $c$, $w_{target}$, $S$, $I$ do
    $e_c \leftarrow$ Text_Encoder($c$); # the complete context representations
    $e^{w_{\text{target}}} \leftarrow e_c$; # the anchor vector based on $w_{\text{target}}$
    # the original unimodal sense/image representations
    $v_{\text{sense}} \leftarrow$ Text_Encoder($S$);
    $v_{\text{image}} \leftarrow$ Image_Encoder($I$);
    # the generated implicit image/sense representations
    $v_{\text{image}}^{\text{sense-gen}} \leftarrow$ SIG($v_{\text{sense}}$);
    $v_{\text{sense}}^{\text{image-gen}} \leftarrow$ ISG($v_{\text{image}}$);
    # the semantically enriched multimodal representations
    $v_{\text{multi}}^{\text{sense-gui}} \leftarrow$ Fusion($v_{\text{sense}}, v_{\text{image}}^{\text{sense-gen}}$);
    $v_{\text{multi}}^{\text{image-gui}} \leftarrow$ Fusion($v_{\text{image}}, v_{\text{sense}}^{\text{image-gen}}$);
    # the anchor-to-sense and anchor-to-image similarities
    $sim(W2S) \leftarrow e^{w_{\text{target}}} \cdot (v_{\text{multi}}^{\text{sense-gui}})^T$;
    $sim(W2I) \leftarrow e^{w_{\text{target}}} \cdot (v_{\text{multi}}^{\text{iamge-gui}})^T$;
    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
    # the pseudo sense/image representations
    $v_{\text{sense}}^{\text{pseudo}} \leftarrow$ Text_Momentum_Encoder($S$);
    $v_{\text{image}}^{\text{pseudo}} \leftarrow$ Image_Momentum_Encoder($I$);
    # the SIG and IGS similarities
    $sim(SIG) \leftarrow v_{\text{image}}^{\text{sense-gen}} \cdot (v_{\text{image}}^{\text{pseudo}})^T$;
    $sim(ISG) \leftarrow v_{\text{sense}}^{\text{image-gen}} \cdot (v_{\text{sense}}^{\text{pseudo}})^T$;
    # the two generation-based loss
    $\mathcal{L}_{SIG} \leftarrow$ CrossEntropyLoss($sim(SIG), labels(SIG)$);
    $\mathcal{L}_{ISG} \leftarrow$ CrossEntropyLoss($sim(ISG), labels(ISG)$);
    # the two understanding-based loss
    $\mathcal{L}_{W2S} \leftarrow$ CrossEntropyLoss($sim(W2S), labels(W2S)$);
    $\mathcal{L}_{W2I} \leftarrow$ CrossEntropyLoss($sim(W2I), labels(W2I)$);
end
```

Four contrastive losses (Hadsell et al., 2006) are defined to optimize four training objectives jointly, comprising two generation-based objectives (SIG loss and ISG loss) and two understanding-based objectives (W2S loss and W2I loss).

The two generation-based objectives make the generated implicit image-text complementary information close to ground truth, to ensure that the enriched multimodal representations are semantically correct. The contrastive loss $\mathcal{L}_{SIG}$ is defined as a cross-entropy $\mathcal{H}$ between the sense-generated image vector $v_{\text{image}}^{\text{sense-gen}}$ and the pseudo-target image vector $v_{\text{image}}^{\text{pseudo}}$:

$$\mathcal{L}_{SIG} = \mathbb{E}_{(I^{\text{gen}}, I^{\text{pse}}) \sim D} \mathcal{H}\left(\mathcal{Y}^{SIG}, \mathcal{P}^{SIG}\right) \tag{11}$$

$\mathcal{Y}$ indicates the ground-truth multi-label one-hot similarity, where negative pairs have a probability of 0 and the positive pairs have a probability of 1. Similarly, the image-generated sense vector $v_{\text{sense}}^{\text{image-gen}}$ and the pseudo-target sense vector $v_{\text{sense}}^{\text{pseudo}}$ are used to calculate $\mathcal{L}_{ISG}$:

$$\mathcal{L}_{ISG} = \mathbb{E}_{(S^{\text{gen}}, S^{\text{pse}}) \sim D} \mathcal{H}\left(\mathcal{Y}^{ISG}, \mathcal{P}^{ISG}\right) \tag{12}$$

The two understanding-based objectives ensure that the PolCLIP model accurately identifies the semantically optimal senses and images. The anchor $e^{w_{\text{target}}}$ and the sense-guided multimodal vector $v_{\text{multi}}^{\text{sense-gui}}$ are used to calculate $\mathcal{L}_{W2S}$:

$$\mathcal{L}_{W2S} = \mathbb{E}_{(W,S) \sim D} \mathcal{H}\left(\mathcal{Y}^{W2S}, \mathcal{P}^{W2S}\right) \tag{13}$$

Also, the anchor $e^{w_{\text{target}}}$ and the image-guided multimodal vector $v_{\text{multi}}^{\text{image-gui}}$ are used to calculate $\mathcal{L}_{W2I}$:

$$\mathcal{L}_{W2I} = \mathbb{E}_{(W,I) \sim D} \mathcal{H}\left(\mathcal{Y}^{W2I}, \mathcal{P}^{W2I}\right) \tag{14}$$

Finally, the full training objective of PolCLIP is:

$$\mathcal{L} = \mathcal{L}_{SIG} + \mathcal{L}_{ISG} + \mathcal{L}_{W2S} + \mathcal{L}_{W2I} \tag{15}$$

In summary, the entire training process is performed under the first objective of the Multimodal-WSD task. The pesudocode of training the PolCLIP model is abstracted in Algorithm 1.

3.3 Inference of the PolCLIP model

The existing WSD test suites suffer from the modality missing issues, resulting in Multimodal-WSD models being evaluated separately under Textual-WSD and Visual-WSD settings. Therefore, the inference process of the PolCLIP model is tested only under the second objective of the Multimodal-WSD task.

To further stimulate the potential of the PolCLIP model in understanding polysemous text, a semantic enhancement is implemented for contexts during the inference procedure. Different to those methods that call APIs, we develop a disambiguation-oriented GPT-3.5 (D-GPT) to generate intended lexical definitions of a word in contexts. Fine-tuning on a random selection of 50,000 data from SemCor, D-GPT is developed based on the gpt-3.5-turbo-1106 model which is one of the latest fine-tunable GPT models released by OpenAI. Each piece of fine-tuning data consists of a query and an output. The query includes a context with a specified ambiguous word. The output is the ground-truth sense of the ambiguous word in the context. The fine-tuning experiment took approximately 175 minutes. More fine-tuning details are provided in Appendix A. After fine-tuning, D-GPT generates lexical definitions based on queries. For example, a query "What does the word 'bank' mean in the context 'They pulled the canoe up on the bank'?" is fed into D-GPT. Then, D-GPT possibly generates an interpretation "The word 'bank' refers to the side or edge of a river, lake, or other body of water".

Algorithm 2: Pseudocode of PolCLIP Inference

```
input  : an augmented context $c$ with an ambiguous target word $w_{\text{target}}$;
         the candidates $Cand$ with senses or images;
output : the ranked candidates $Cand_{\text{ranked}}$;
         the semantically optimal sense or image $O_{\text{best}}$;
$e_c \leftarrow$ Text_Encoder($c$); # the complete context representations
$e^{w_{\text{target}}} \leftarrow e_c$; # the anchor vector based on $w_{\text{target}}$
if only senses in $Cand$ then
    # for Textual-WSD
    $v_{\text{sense}} \leftarrow$ Text_Encoder($Cand$);
    # the sense-generated image representations
    $v_{\text{image}}^{\text{sense-gen}} \leftarrow$ SIG($v_{\text{sense}}$);
    # the sense-guided multimodal representations
    $v_{\text{multi}} \leftarrow$ Fusion($v_{\text{sense}}$ , $v_{\text{image}}^{\text{sense-gen}}$);
else
    # for Visual-WSD
    $v_{\text{image}} \leftarrow$ Image_Encoder($Cand$);
    # the image-generated sense representations
    $v_{\text{sense}}^{\text{image-gen}} \leftarrow$ ISG($v_{\text{image}}$);
    # the image-guided multimodal representations
    $v_{\text{multi}} \leftarrow$ Fusion($v_{\text{image}}$ , $v_{\text{sense}}^{\text{image-gen}}$);
end
$similarity \leftarrow e^{w_{\text{target}}} \cdot (v_{\text{multi}})^T$;
$Cand_{\text{ranked}} \leftarrow top_k(similarity)$; # $k$ is the number of candidates
$O_{\text{best}} \leftarrow \arg\max(Cand_{\text{ranked}})$; # the semantically optimal sense or image
```

During the testing phase, all the original contexts in WSD test sets are concatenated with the lexical definitions generated by D-GPT, to create semantically augmented contexts. For instance, an augmented context is "They pulled the canoe up on the bank. The word 'bank' refers to the side or edge of a river, lake, or other body of water". These augmented contexts are subsequently fed into the PolCLIP model. The trained SIG and ISG modules support the PolCLIP model to address Multimodal-WSD even when any modality is missing. The inference procedure of the PolCLIP model is abstracted in Algorithm 2.

3.4 Training data

The PolCLIP model relies on large-scale aligned image-sense pairs to learn multimodal polysemy knowledge. Thus, we construct a disambiguation-oriented image-sense dataset by integrating SemCor (Miller et al., 1993) and VWSD-KB (Yang et al., 2023), to achieve the training objective of unified image-text WSD. SemCor is the most prevalent dataset for training Textual-WSD models. VWSD-KB contains multimodal data such as words, senses, and images. The offline BabelNet v5.2 and BabelPic are used to collect relevant images to the senses in SemCor and VWSD-KB. The detailed construction process is provided in Appendix B.

After construction, all word senses in SemCor and VWSD-KB are aligned with at least one image and at most five images. We filter out the collected images that are pornographic, violent, or invalid and conduct a manual validation to ensure there is no data leakage. The SemCor associated with images is called Image-Enhanced SemCor. Table 1 displays the statistical details of Image-Enhanced SemCor and VWSD-KB, comprising a total of 274,505 English instances (each instance includes a context with at least one ambiguous target word). An example of the disambiguation-oriented image-sense dataset is shown in Figure 1. Given a context "He had seen the Andromeda tree in Japan", there are four candidate senses for the ambiguous target word "Andromeda", and each sense corresponds to five images. The most semantically relevant sense and images to this context are **Sense 4** and its five associated images.

Item types	Image-Enhanced SemCor	VWSD-KB
# of instances	226,036	48,469
# of ambiguous target words	33,657	24,989
# of senses	39,201	31,306
# of images	181,123	111,575

Table 1: The statistical details of Image-Enhanced SemCor and VWSD-KB.

4 Experiments and Results

4.1 Datasets

Due to the training objective of unified image-text WSD, we opt not to use the validation sets from Textual-WSD or Visual-WSD. We allocate 80% of the combined Image-Enhanced SemCor and VWSD-KB datasets as the training set and reserve the remaining 20% as the validation set. XL-WSD (Pasini et al., 2021), an extra-large evaluation framework for Textual-WSD, is employed to evaluate the model performance on Textual-WSD. XL-WSD is widely used since it encompasses six English all-words Textual-WSD benchmark datasets, including SensEval-2 (SE2, (Palmer et al., 2001)), SensEval-3 (SE3, (Snyder and Palmer, 2004)), SemEval-2007 (SE7, (Navigli et al., 2007)), SemEval-2010 (SE10, (Agirre et al., 2010)), SemEval-2013 (SE13, (Navigli et al., 2013)), and SemEval-2015 (SE15, (Moro and Navigli, 2015)). These six benchmark datasets comprise a total of 8,517 English instances for testing. SemEval-2023 (SE23, (Raganato et al., 2023)) is used to assess the model performance on Visual-WSD, as it is currently the most widely used Visual-WSD benchmark containing 463 English instances.

Training Data	Models	Textual-WSD							Visual-WSD	
		SE2	SE3	SE7	SE10	SE13	SE15	ALL	SE23	
		F1-score(%)							HR@1(%)	MRR@10(%)
Zero-shot	Openai/CLIP-ViT-L/14	**53.07**	**47.19**	35.60	37.50	57.18	**53.52**	**49.40**	57.45	72.60
	Laion/CLIP-ViT-L/14	49.27	43.88	31.80	33.70	53.38	49.96	45.73	56.87	70.28
	Laion/CLIP-ViT-H/14	51.47	46.45	**36.52**	**39.10**	**58.78**	51.92	49.21	60.70	75.68
	UVWSD	-	-	-	-	-	-	-	**80.50**	**87.60**
Image-Enhanced SemCor	Openai/CLIP-ViT-L/14	70.01	66.48	61.35	71.96	72.35	69.11	69.41	76.67	84.20
	Laion/CLIP-ViT-L/14	71.01	65.96	62.34	72.92	71.44	68.37	69.50	75.38	84.02
	Laion/CLIP-ViT-H/14	71.00	69.63	65.24	74.53	75.36	71.61	71.83	77.04	84.37
	BEM	78.29	75.96	66.64	80.71	81.38	81.72	78.53	-	-
	SACE	80.29	78.67	70.57	82.71	80.86	83.73	80.30	-	-
	Z-Reweighting	79.98	77.04	67.72	82.01	79.94	82.81	79.32	-	-
	FCLL	-	-	-	-	-	-	-	80.13	87.41
	PolCLIP$_{base}$	82.22	79.89	70.56	85.22	82.79	85.66	82.06	79.48	85.00
	PolCLIP$_{base}$ with D-GPT	**83.74**	**81.41**	**72.09**	**86.16**	**84.31**	**87.18**	**83.49**	**82.94**	**88.55**
Image-Enhanced SemCor + VWSD-KB	Openai/CLIP-ViT-L/14	71.90	68.23	63.21	73.76	74.12	67.80	70.85	75.98	83.93
	Laion/CLIP-ViT-L/14	69.86	67.37	60.46	72.85	73.24	70.00	69.93	77.88	84.68
	Laion/CLIP-ViT-H/14	73.90	70.37	65.24	75.85	76.24	70.80	73.04	77.46	84.60
	BEM	78.09	75.03	67.65	80.01	78.45	80.09	77.46	-	-
	SACE	81.93	79.71	71.71	84.35	79.22	84.87	81.09	-	-
	Z-Reweighting	79.53	77.03	68.92	81.07	82.50	82.09	79.53	-	-
	FCLL	-	-	-	-	-	-	-	81.37	87.69
	PolCLIP$_{large}$	82.76	82.39	71.11	**85.18**	**85.29**	86.20	82.60	82.28	87.98
	PolCLIP$_{large}$ with D-GPT	**84.66**	**82.43**	**72.97**	83.39	84.91	**86.40**	**83.62**	**83.59**	**90.07**

Table 2: Comparison with state-of-the-art methods on WSD benchmark test sets. Bold numbers indicate results of the SOTA model, and underlined numbers denote results of the second best model.

4.2 Implementation details

Settings. Our model is implemented on Pytorch 2.0.1 and 4 RTX 4090 GPUs. Both the text encoder and image encoder are initialized by CLIP-ViT-L/14 (Radford et al., 2021). All parameters of the text encoder are optimized to preserve the maximum capability of the model to process ambiguous knowledge at the text level, while the image encoder is completely frozen to reduce the computational cost. The sense batch size is set to 50, the image batch size is set to 250. Following ALBEF (Li et al., 2021), the momentum is set to 0.005. AdamW is applied to optimize model parameters with a learning rate of 1e-04 and weight decay of 0.05. The image resolution is specified as 224×224, and the maximum text length is set to 77. The training process of the PolCLIP model requires approximately 2.5 GPU hours per epoch, with a total of 20 epochs. F1-score is used to evaluate the model performance on Textual-WSD. Hit Rate at 1 (HR@1, i.e., accuracy) and Mean Reciprocal Rank at 10 (MRR@10) are used to assess the model performance on Visual-WSD.

Baselines. We train PolCLIP$_{base}$ using Image-Enhanced SemCor and PolCLIP$_{large}$ using the combination of Image-Enhanced SemCor and VWSD-KB. In the testing phase, both PolCLIP$_{base}$ and PolCLIP$_{large}$ are integrated with D-GPT for semantic enhancement. Our model is compared with recent state-of-the-art (SOTA) methods including (1) SOTA models in Textual-WSD: BEM (Blevins and Zettlemoyer, 2020), SACE (Wang and Wang, 2021) and Z-Reweighting (Su et al., 2022), (2) SOTA models in Visual-WSD: FCLL (Yang et al., 2023) and UVWSD (Kwon et al., 2023), (3) SOTA models in image-text learning tasks: Openai/CLIP-VIT-L/14 (Radford et al., 2021), Laion/CLIP-VIT-L/14 and Openai/CLIP-VIT-H/14 (Schuhmann et al., 2022). More details about baseline models are provided in Appendix C. For a fair comparison, these baseline models are retrained using Image-Enhanced SemCor and VWSD-KB.

4.3 WSD results

The comparison results between PolCLIP and the baseline models on WSD benchmark test sets are shown in Table 2. The PolCLIP model achieves the state-of-the-art performance. SACE and FCLL are the second best models for Textual-WSD and Visual-WSD, respectively. Without D-GPT, PolCLIP$_{large}$ reaches an F1-score of 82.60% on all the Textual-WSD test data, which is 1.51% higher than SACE. It attains an HR@1 of 82.28% and a MRR@10 of 87.98% on Visual-WSD, which are 0.91% and 0.29% higher than FCLL respectively. With D-GPT, the performance of both PolCLIP$_{base}$ and PolCLIP$_{large}$ is enhanced thanks to the semantically augmented contexts with lexical definitions.

In this situation, PolCLIP$_{large}$ gains an F1-score of 83.62% on Textual-WSD, which is 2.53% higher than SACE. It attains an HR@1 of 83.59% and a MRR@10 of 90.07% on Visual-WSD, which are 2.22% and 2.38% higher than FCLL respectively.

Specifically, without training, Openai/CLIP-VIT-L/14, Laion/CLIP-VIT-L/14, and Openai/CLIP-VIT-H/14 (collectively called CLIPs) obtain zero-shot F1-scores below 50% on all the Textual-WSD test data, due to the pre-training data and goal of CLIPs do not target WSD. Conversely, UVWSD obtains over 80% zero-shot HR@1 on Visual-WSD. While only using the Image-Enhanced SemCor as the training set, PolCLIP$_{base}$ outperforms CLIPs, BEM, SACE, and Z-Reweighting on Textual-WSD, even though its performance on Visual-WSD is slightly inferior to FCLL. When the combination of Image-Enhanced SemCor and VWSD-KB is used as the training set, PolCLIP$_{large}$ shows further improvement over PolCLIP$_{base}$ and surpasses all baseline models. Additionally, the effectiveness of the disambiguation-oriented image-sense dataset is proven, with the performance of all the baseline models on WSD benchmarks being bolstered.

4.4 Ablation study

An ablation study is conducted to reveal the contribution of each module and the results are reported in Table 3. For the two generation-based training objectives, the Sense-to-Image Generation module is removed first, which corresponds to $\mathcal{L}_{SIG}$. In this scenario, the F1-score of PolCLIP$_{large}$ on Textual-WSD decreases by 8.08% and the HR@1 on Visual-WSD drops by 1.07%. This indicates that the implicit image information generated by the SIG module aids PolCLIP$_{large}$ in acquiring enriched multimodal representations. Secondly, the Image-to-Sense Generation module is removed, which corresponds to $\mathcal{L}_{ISG}$. The HR@1 of PolCLIP$_{large}$ on Visual-WSD decreases by 5.82% and the F1-score on Textual-WSD drops by 1.43%. This demonstrates that the implicit sense information generated by the ISG module also facilitates PolCLIP$_{large}$ obtaining deep polysemy knowledge.

Models	Textual-WSD ALL F1-score (%)	Visual-WSD HR@1 (%)
w/o-SIG	74.52 (-8.08)	81.21 (-1.07)
w/o-ISG	81.17 (-1.43)	76.46 (-5.82)
w/o-W2I	82.95 (+0.35)	9.29 (-72.99)
w/o-W2S	19.73 (-62.87)	82.72 (+0.44)

Table 3: Ablation study of PolCLIP$_{large}$ on the WSD benchmark test sets.

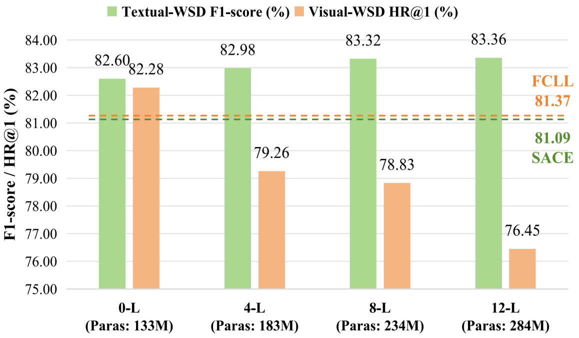


Figure 3: The experimental results of PolCLIP$_{large}$ with different layer numbers for optimizing the image encoder.

Regarding the two understanding-based training objectives, the alignment process between the anchor focused on ambiguous target words and candidate images is eliminated first, which corresponds to $\mathcal{L}_{W2I}$. This means that PolCLIP$_{large}$ exclusively trains for Textual-WSD. At this point, the HR@1 of PolCLIP$_{large}$ on Visual-WSD is only 9.29%. PolCLIP$_{large}$ is regarded as a model that makes random selections for Visual-WSD, since it is required to choose one from ten candidate images for each instance. However, the trade-off problem caused by multi-task training objectives allows its F1-score on Textual-WSD has a 0.35% improvement. Secondly, the alignment process between the anchor and candidate senses is eliminated, corresponding to $\mathcal{L}_{W2S}$. This means that PolCLIP$_{large}$ exclusively trains for Visual-WSD. In this situation, the F1-score of PolCLIP$_{large}$ on Textual-WSD is only 19.73% but the HR@1 on Visual-WSD increases by 0.44%. The explanations are similar. An additional experiment, which investegates the generality of D-GPT for WSD, is provided in Appendix D.

Overall, the two generation-based modules actually facilitate PolCLIP acquiring multimodal polysemy knowledge. The two understanding-based alignment processes are the most critical components, since they maximize the similarities between contexts and senses/images in a feature space.

4.5 Optimal layer number for optimizing image encoder

To reduce the computational cost, we opt not to optimize all parameters of the image encoder. With all parameters of the text encoder being optimized, the last 0/4/8/12 layers of the image encoder are separately optimized to investigate their impact on the

Models	NOUN	VERB	ADJ	ADV
SACE	82.84	74.23	84.77	81.90
PolCLIP$_{large}$	84.23 (+1.39)	74.38 (+0.15)	88.13 (+3.36)	87.62 (+5.72)

Table 4: The F1-scores of SACE and PolCLIP$_{large}$ for ambiguous target words with different parts of speech.

model performance. The results of PolCLIP$_{large}$ with different layer numbers for optimizing the image encoder are displayed in Figure 3. When zero layers are optimized (meaning the image encoder is completely frozen), PolCLIP$_{large}$ has the smallest parameter size and the SOTA results on WSD benchmarks. Therefore, this model configuration is selected as our best model (as reported in Table 2). It is interesting that when more layers are optimized, the model performance gradually improves in a small way for Textual-WSD, but drops significantly for Visual-WSD. This is contrary to our expectations. Theoretically, optimizing more layers of the image encoder should enhance the model ability to capture image knowledge. We speculate that redundant knowledge, introduced by some noisy image-sense pairs in the training set, increases the model's training burden. Thus, refining this disambiguation-oriented image-sense dataset would be valuable.

4.6 Analysis on model performance for different PoS

Since some words may present different parts of speech (PoS) in contexts, exploring the model performance for ambiguous target words with different PoS is beneficial to reveal the unique advantages of the PolCLIP model. The F1-score results of SACE and PolCLIP$_{large}$, trained on the combination of Image-Enhanced SemCor and VWSD-KB, are shown in Table 4. Compared with SACE, PolCLIP$_{large}$ has improvements of 1.39%, 0.15%, 3.36% and 5.72% in NOUN, VERB, ADJ and ADV respectively. One of the challenges in WSD tasks is the difficulty of understanding non-concrete words accurately, which are often adjectives or adverbs. PolCLIP$_{large}$ happens to have a more obvious improvement in adjectives and adverbs. Therefore, we believe that PolCLIP$_{large}$ has favorable adaptability and flexibility for ambiguous target words with different PoS, due to its generation-based advantages. On the one hand, it can supplement tangible senses or images with semantically concrete images or descriptions. On the other hand, it can also supplement non-concrete senses or images with semantically abstract images or descriptions. The case study and visualizations of the implicit image-text complementary information generated by the SIG and ISG modules for concrete and non-concrete examples are provided in Appendix E.

5 Conclusion

This paper proposes a unified image-text WSD model PolCLIP, which achieves state-of-the-art performance on Textual-WSD and Visual-WSD benchmark datasets. Extensive experimental results prove the effectiveness of our image-text complementarity strategy. A series of in-depth explorations of the model architecture demonstrate the Sense-to-Image Generation module and the Image-to-Sense Generation module can generate effective multimodal complementary representations. The disambiguation-oriented image-sense dataset empirically facilitates WSD models understanding of multimodal polysemy knowledge. This may provide a benchmark for the future Multimodal-WSD task. Source codes and datasets are publicly released at: https://github.com/CharlesYang030/PolCLIP.

Limitation

In this work, we did not explore the Multimodal-WSD models on multilingual data. The proposed image-text complementarity strategy was employed to address the modality missing issues in unimodal WSD datasets. However, the PolCLIP model still lacked the ability to generate realistic senses and images that can be intuitively validated in terms of their semantics. Moreover, D-GPT was built upon GPT-3.5 which was a closed-source model and may be changed, updated, or even discontinued. In future work, we plan to further expand the disambiguation-oriented image-sense dataset to cover more languages. We will also develop a large generic model suitable for Multimodal-WSD.

Acknowledgements

The work is supported by grants from National Natural Science Foundation of China (62372189), the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E09/22), and the Scientific Research Innovation Project of Graduate School of South China Normal University (2024KYLX090).

References

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 75–80, Uppsala, Sweden. Association for Computational Linguistics.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. 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.

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.

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

Agostina Calabrese, Michele Bevilacqua, Roberto Navigli, et al. 2020b. Evilbert: Learning task-agnostic multimodal sense embeddings. In *IJCAI*, pages 481–487. International Joint Conferences on Artificial Intelligence Organization.

Guanhua Chen, Qiqi Xu, Choujun Zhan, Fu Lee Wang, Kai Liu, Hai Liu, and Tianyong Hao. 2024. Improving open intent detection via triplet-contrastive learning and adaptive boundary. *IEEE Transactions on Consumer Electronics*, 70:2806–2816.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275, Online. Association for Computational Linguistics.

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 (SemEval-2023)*, pages 155–162, Toronto, Canada. Association for Computational Linguistics.

Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation. 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 1998–2004, Minneapolis, Minnesota. Association for Computational Linguistics.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–192, San Diego, California. Association for Computational Linguistics.

Omid Ghahroodi, Seyed Arshan Dalili, Sahel Mesforoush, and Ehsaneddin Asgari. 2023. 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)*, pages 2160–2163, Toronto, Canada. Association for Computational Linguistics.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Anastasia Kritharoula, Maria Lymperaiou, and Giorgos Stamou. 2023. Large language models and multimodal retrieval for visual word sense disambiguation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13053–13077, Singapore. Association for Computational Linguistics.

Sunjae Kwon, Rishabh Garodia, Minhwa Lee, Zhichao Yang, and Hong Yu. 2023. Vision meets definitions: Unsupervised visual word sense disambiguation incorporating gloss information. In *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1583–1598, Toronto, Canada. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3534–3540, Hong Kong, China. Association for Computational Linguistics.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Di Mo, Bangrui Huang, Haitao Wang, Xinyu Cao, Keqin Gan, Jie Wei, Heng Weng, and Tianyong Hao. 2023. Sclert: A span-based joint model for measurable quantitative information extraction from chinese texts. *IEEE Transactions on Consumer Electronics*, 70:3361–3371.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of babelnet: A survey. In *IJCAI*, pages 4559–4567.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24, Toulouse, France. Association for Computational Linguistics.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual 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, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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)*, pages 2227–2234, Toronto, Canada. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT

2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681, Dubrovnik, Croatia. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–46, Online. Association for Computational Linguistics.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

Ying Su, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. Rare and zero-shot word sense disambiguation using Z-Reweighting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4713–4723, Dublin, Ireland. Association for Computational Linguistics.

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.

Ming Wang and Yinglin Wang. 2021. Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5218–5229, 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)*, pages 506–511, Toronto, Canada. Association for Computational Linguistics.

Qihao Yang, Xuelin Wang, Yong Li, Lap-Kei Lee, Fu Lee Wang, and Tianyong Hao. 2024. Mta: A lightweight multilingual text alignment model for cross-language visual word sense disambiguation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12166–12170. IEEE.

Xudong Zhang, Tiange Zhen, Jing Zhang, Yujin Wang, and Song Liu. 2023. 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)*, pages 439–446, Toronto, Canada. Association for Computational Linguistics.

A Fine-tuning D-GPT

A fine-tuned disambiguation-oriented GPT-3.5 (D-GPT) is developed to generate lexical definitions for ambiguous target words during the testing phase. The gpt-3.5-turbo-1106 model is chosen as the backbone model, since it is one of the latest fine-tunable GPT models released by OpenAI and outperforms several open-source LLMs with smaller parameter sizes in terms of inference capabilities. Constrained by fine-tuning costs, we randomly collect 50,000 data from SemCor to serve as the fine-tuning dataset. Each piece of fine-tuning data consists of a query and an output. The OpenAI fine-tuning platform requires users to simply package their fine-tuning corpus into message-style data and upload it for end-to-end fine-tuning. Following OpenAI's guidelines[1], the format of message-style data for fine-tuning D-GPT is shown in Figure 4. After fine-tuning, D-GPT generates lexical definitions based on queries. For example, a query "What does the word 'bank' mean in the context 'They pulled the canoe up on the bank'?" is fed into D-GPT. Then, D-GPT possibly generates an interpretation "The word 'bank' refers to the side or edge of a river, lake, or other body of water". Even though the output of D-GPT seems to be the answer to generative WSD tasks, we aim to use the output to build a semantically augmented context because WSD is defined as a classification task in this work.

```
"messages": [
{"role": "system", "content": "I am a   factual
          chatbot that is excellent in word sense
          disambiguation."},
{"role": "user", "content": "What does the
          word W mean in the context C ?"},
{"role": "assistant", "content":  GT}]
```

Figure 4: The format of message-style data for fine-tuning D-GPT. The red GT denotes the ground-truth sense.

B Construction of the disambiguation-oriented image-sense datasets

The disambiguation-oriented image-sense dataset was constructed based on SemCor (Miller et al., 1993) and VWSD-KB (Yang et al., 2023) datasets. Specifically, the offline version of BabelNet[2] v5.2 was employed to collect a list of relevant images based on each sense in these two datasets. If there were more than five available images in the list, the top five were selected; otherwise, all images were retained. However, a minority of the senses failed to associate with any image through BabelNet, as they were typically non-concrete, like expressing sadness. Thus, BabelPic (Calabrese et al., 2020a), an image-text dataset for non-concrete concepts, was utilized to find images for a part of non-concrete senses based on babel-ids. Furthermore, those senses that we had collected relevant images were set as an internal knowledge base. For each sense s_i that was not included in either BabelNet or BabelPic, RoBERTa[3] was used to identify the three senses most semantically similar to s_i within this internal knowledge base, based on text similarity. The first image from each of these three most similar senses was aggregated as the set of images corresponding to s_i. The entire construction process enabled all word senses in SemCor and VWSD-KB to be aligned with at least one image and at most five images.

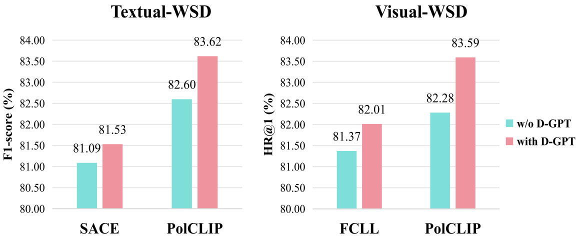


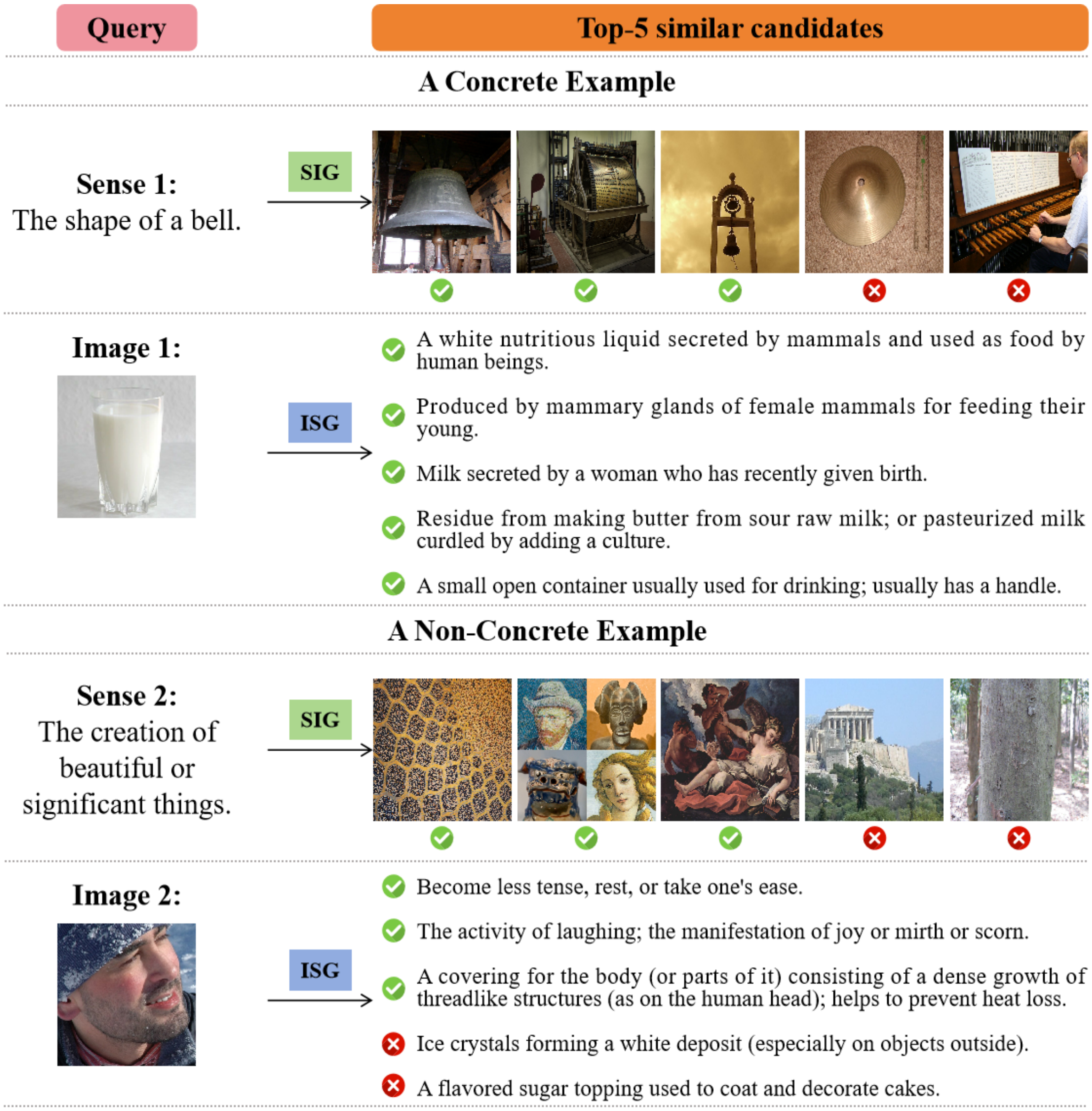
Figure 5: The evaluation results of SACE, FCLL, and PolCLIP$_{large}$ on the benchmark test sets after integrating D-GPT.

After construction, we had conducted a comparative analysis between the proposed training dataset and the existing benchmarks. For Textual-WSD, we examined all contexts in our training dataset and the test set, ensuring there were no duplicates. For Visual-WSD, all images were uniformly transformed into 3×224×224 tensors, and then paired with their respective contexts to form <context, tensor> pairs (Mo et al., 2023; Chen et al., 2024). There were no duplicate <context, tensor> pairs. Through this manual verification, we ensured there was no data leakage.

[1] The fine-tuning guidelines of OpenAI: https://platform.openai.com/docs/guides/fine-tuning

[2] BabelNet (updated to version 5.3): https://babelnet.org/

[3] The hugging face link of RoBERTa: https://huggingface.co/FacebookAI/roberta-large



✓ A white nutritious liquid secreted by mammals and used as food by human beings.

✓ Produced by mammary glands of female mammals for feeding their young.

✓ Milk secreted by a woman who has recently given birth.

✓ Residue from making butter from sour raw milk; or pasteurized milk curdled by adding a culture.

✓ A small open container usually used for drinking; usually has a handle.

✓ Become less tense, rest, or take one's ease.

✓ The activity of laughing; the manifestation of joy or mirth or scorn.

✓ A covering for the body (or parts of it) consisting of a dense growth of threadlike structures (as on the human head); helps to prevent heat loss.

✗ Ice crystals forming a white deposit (especially on objects outside).

✗ A flavored sugar topping used to coat and decorate cakes.

Figure 6: Visualizations of the implicit image-text complementary information generated by the SIG and ISG modules for concrete and non-concrete examples.

C Baselines

The details of baselines are as follows:

SOTA models in Textual-WSD: BEM (Blevins and Zettlemoyer, 2020) adopted two text encoders and focuses on the representations of ambiguous target words rather than the complete context representations. SACE (Wang and Wang, 2021) employed an interactive context exploitation method and selects similar sentences from the same document to enhance context representations. Z-Reweighting (Su et al., 2022) utilized a strategy for adjusting training on imbalanced datasets at the word level. These three models obtain outstanding performance on Textual-WSD benchmarks when training exclusively on SemCor.

SOTA models in Visual-WSD: FCLL (Yang et al., 2023) employed a fine-grained image-text contrastive learning mechanism and benefited from VWSD-KB. It won first place in SemEval-2023 Task 1. UVWSD (Kwon et al., 2023) did not necessitate training but achieved remarkable performance by employing Bayesian inference to incorporate sense definitions.

SOTA models in image-text learning tasks: Openai/CLIP-VIT-L/14 (Radford et al., 2021), Laion/CLIP-VIT-L/14 and Openai/CLIP-VIT-H/14 (Schuhmann et al., 2022) all employ a dual-stream architecture to learn image-text knowledge, similar to our PolCLIP model. The first model is pre-trained on over 400 million image-text pairs, and the latter two are pre-trained on the English subset of LAION-5B (Schuhmann et al., 2022), a large publicly available image-text dataset.

D Generality of D-GPT

An additional experiment is conducted to explore the generality of D-GPT for WSD tasks. Specifically, D-GPT is integrated with SACE and FCLL, and its impact on the performance of these two models and PolCLIP$_{large}$ is illustrated in Figure 5. D-GPT indeed enhances the performance of these three WSD models. This indicates that using lexical definitions generated by D-GPT to create semantically augmented contexts is a general-purpose and convenient pipeline for WSD tasks. It can be applied to various WSD models. Furthermore, compared to the evaluation results without D-GPT, PolCLIP$_{large}$ shows an improvement of 1.02% F1-score and 1.31% HR@1. These are respectively higher than the 0.44% F1-score increase of SACE on Textual-WSD and the 0.64% HR@1 increase of FCLL on Visual-WSD. This also leads us to believe that PolCLIP could gain more when dealing with contexts that are semantically more accurate, thanks to the image-text complementarity strategy.

E Effectiveness of the SIG and ISG modules

In order to further intuitively reveal the effectiveness and importance of the two generation-based modules (i.e., SIG and ISG), the generated implicit image-text complementary information is visualized. Two groups of concrete and non-concrete examples are collected from the test set. Each group of examples contains a sense and an image. They are fed into the trained SIG and ISG modules respectively, and then a sense-generated image vector and an image-generated sense vector are output. All senses and images in the training set are transformed into vectors by text and image encoders, serving as two separate candidate pools. By calculating vector similarity, the top-5 most similar images can be identified from the image candidate pool based on the sense-generated image vector. Also, the top-5 most similar senses can be identified from the sense candidate pool based on the image-generated sense vector. Visualizations of these two groups of concrete and non-concrete examples are shown in Figure 6. For the concrete example, the top-3 images are semantically consistent with **Sense 1**. Even if the last two images do not depict the shape of a bell, they are related to music or sound, which is one of the functions of bells. Based on **Image 1** related to milk, the retrieved five senses are all semantically correct. For the non-concrete example, the top-3 images are semantically relevant to **Sense 2** related to beauty. Even based on **Image 2**, which primarily shows a man's face expressing pleasure, the top-3 senses accurately capture concepts of relaxation, laughter, and beard.

In summary, the SIG and ISG modules are reliable in generating effective multimodal complementary information.