From Sparse to Sense-Grounded: Wikipedia Training for Ukrainian Visual-WSD

Yurii Laba
Ukrainian Catholic University
Lviv, Ukraine
laba@ucu.edu.ua

Rostyslav O. Hryniv
Ukrainian Catholic University
Lviv, Ukraine
rhryniv@ucu.edu.ua

Abstract

Visual Word Sense Disambiguation (Visual-WSD) requires ranking the correct image for an ambiguous word given a short trigger phrase. For low-resource languages, it is bottlenecked by scarce sense-level benchmarks and limited sense-aligned multimodal supervision. We study Ukrainian and (i) extend the Ukrainian Visual-WSD benchmark from 87 to 381 instances and benchmark multilingual CLIP checkpoints and multimodal large models, and (ii) introduce two scalable Wikipedia-derived dataset construction methods. Using compute-efficient adaptation we fine-tune a multilingual CLIP backbone and show that sense-grounded supervision drives the improvements: combining our two Wikipedia-derived datasets improves HIT@1 from 37.00% to 43.05%.

1 Introduction

Many words are lexically ambiguous: the same surface form can express distinct senses, and correct interpretation depends on context. Word Sense Disambiguation (WSD) is therefore central to language understanding and downstream tasks such as translation and retrieval. In practice, context is often multimodal: visual evidence can directly resolve meaning (e.g., whether crane refers to the bird or the machine). This motivates Visual Word Sense Disambiguation (Visual-WSD), where a model selects the image that matches an intended sense given minimal textual context.

Visual-WSD was introduced in SemEval-2023 as a multilingual retrieval task: given an ambiguous word and a short trigger phrase, models rank ten candidate images (Raganato et al., 2023). The task provides a controlled test of whether vision-language models capture fine-grained sense distinctions rather than relying on coarse semantic similarity.

Extending Visual-WSD to low-resource languages faces two coupled bottlenecks. First, evaluation requires curated sense-level benchmarks. Second, adaptation requires multimodal supervision that is both in-language and aligned with sense distinctions. Ukrainian highlights both: off-the-shelf multilingual vision-language models lag on Ukrainian Visual-WSD compared to English baseline (Laba et al., 2024), while high-quality Ukrainian image-text supervision is scarce. This motivates leveraging the few Ukrainian-native multimodal sources available at scale.

Wikipedia is one of the few scalable sources of Ukrainian-native multimodal content with implicit sense structure (separate pages for distinct meanings), but its supervision is sparse. In our Ukrainian dump, only 64,130 of 150,000 images include Ukrainian captions; many pages are weakly illustrated and captions are frequently missing. This sparsity forces training text to come from titles, lead definitions, or templatic metadata, creating a mismatch with the short trigger-phrase queries used in Visual-WSD evaluation and making direct contrastive tuning unreliable.

In our study, we extend the Ukrainian Visual-WSD benchmark from 87 to 381 instances[1] and benchmark multilingual CLIP checkpoints and multimodal LMMs on the extended set. We further create two scalable Wikipedia-based sources of sense-aligned supervision (SenseWiki-UA and RA-Wiki-UA), and study compute-efficient adaptation of CLIP-like model via vision-frozen, top-k text-layer tuning, including the trade-off between Visual-WSD gains and general image-to-text/text-to-image retrieval robustness.

2 Related Work

2.1 From Textual WSD to Visual-WSD

Word Sense Disambiguation (WSD) has traditionally relied on textual context and lexical databases

[1]Visual tool to explore the benchmark: https://v-wsd-backend-ac95w.ondigitalocean.app/

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 501–514

such as WordNet, with early benchmarks established through Senseval (Edmonds and Cotton, 2001) and SemEval (Navigli et al., 2013) shared tasks. More recent research extends WSD into the multimodal domain by incorporating visual context, leading to the formulation of Visual Word Sense Disambiguation (Visual-WSD).

2.2 SemEval Visual-WSD and Common Solution Patterns

SemEval-2023 Task 1 introduced Visual-WSD task, where a model selects the correct image (out of ten candidates) for an ambiguous target word given minimal textual context, with datasets in English, Italian, and Farsi and a strong zero-shot CLIP baseline. The task attracted 96 submissions, and 40 systems surpassed the baseline, showing that task-specific heuristics and lightweight augmentation can substantially improve CLIP-style matching (Raganato et al., 2023; Radford et al., 2021).

Across submissions, common strategies include: (i) enriching the short context with lexical or encyclopedic resources (e.g., BabelNet/-WordNet glosses, Wikipedia retrieval) followed by re-ranking (Yang et al., 2023; Dadas, 2023; Ogezi et al., 2023); (ii) expanding queries with LM-generated paraphrases or explanations to better match CLIP's text encoder (Li et al., 2023; Ogezi et al., 2023); and (iii) refining candidate images via captioning or text-to-image generation (e.g., Stable Diffusion) to increase the likelihood of a strong sense match (Mijatovic et al., 2023; Li et al., 2023).

2.3 Low-Resource Visual-WSD and the Ukrainian Setting

The dominant SemEval patterns above implicitly assume (i) broad image-text supervision and/or (ii) high-coverage lexical resources for query expansion and gloss-based disambiguation. For Ukrainian, both are weaker: large-scale Ukrainian image-text supervision is limited, and available lexical resources are often derived from Wikipedia-linked mappings with sparse synonymy compared to high-resource WordNets, which constrains SemEval-style synonym/gloss expansion (Romanyshyn et al., 2024; Laba et al., 2024).

LLMs offer a flexible substitute by generating definitions, paraphrases, or disambiguating contexts, and recent work shows strong WSD capability in high-resource settings (Meconi et al., 2025). However, reliability can degrade in low-resource languages, and translation-mediated prompting is not uniformly robust across languages and domains (Dey et al., 2024). In deployment, inference cost and latency further motivate compute-efficient alternatives, such as adapting CLIP-style retrieval with small amounts of targeted supervision (Zheng et al., 2025).

2.4 Multilingual CLIP Adaptations

Given these constraints, contrastive image-text models like CLIP remain appealing due to their lower inference cost. Since CLIP was originally trained on English data, adapting it for other languages is necessary, and several multilingual adaptations have been proposed.

M-CLIP (Carlsson et al., 2022) replaces CLIP's English text encoder with language-specific alternatives via a cross-lingual teacher-student framework, using translation and alignment without extra image data. AltCLIP (Chen et al., 2023) incorporates a multilingual Transformer (XLM-R) as the text encoder, fine-tuned via distillation from English CLIP and contrastive training on translated pairs. OpenCLIP (Ilharco et al., 2021), leveraging huge multimodal datasets like LAION-5B (Schuhmann et al., 2022), shows improved zero-shot performance for non-English languages. Despite these advancements, multilingual CLIP models still struggle in low-resource settings. For instance, on the Ukrainian Visual-WSD benchmark (Laba et al., 2024), they underperform compared to the English-only CLIP baseline on the English Visual-WSD dataset, highlighting the limits of cross-lingual transfer when Ukrainian supervision is scarce.

2.5 Compute-efficient adaptation of contrastive vision-language models

Contrastive vision-language pre-training on image-text pairs enables strong zero-shot transfer for recognition and retrieval, with CLIP (Radford et al., 2021) as a standard representative. However, full contrastive pre-training is resource-intensive, motivating approaches that adapt pre-trained backbones instead of training from scratch. Locked-image tuning (LiT) freezes the vision encoder and updates the text encoder to improve efficiency while preserving visual representations (Zhai et al., 2022). Initializing the text encoder from a pre-trained language model can further preserve linguistic structure while aligning to visual concepts during tuning (Kim et al., 2024). Beyond full-tower tuning, selective layer updates (e.g., tuning only upper layers) have been explored to balance adaptation and sta-

bility (Zhu et al., 2024). Finally, contrastive tuning often benefits from augmentation on both modalities: image transformations expand visual coverage (Kumar et al., 2024; Yang et al., 2022), while text diversification methods such as back-translation or paraphrasing can increase caption variety without changing meaning (Kim et al., 2024; Li et al., 2024; Chai et al., 2025).

In summary, while multilingual CLIP variants and parameter-efficient tuning methods exist, their application to Visual-WSD is not well understood. There is limited evidence on how to effectively adapt such models for sense-level discrimination, and few scalable methods for constructing sense-grounded image–text supervision. These limitations are especially acute in low-resource languages, where both lexical resources and multimodal data are sparse. Addressing these gaps requires both compute-efficient adaptation strategies and practical pipelines for mining sense-aligned supervision which we propose in this work.

3 Benchmark: Ukrainian Visual-WSD Extension

To enable systematic evaluation of multilingual and multimodal models in low-resource settings, we extend the Ukrainian Visual-WSD benchmark introduced by Laba et al. (2024). The prior release contained 87 benchmark entries. Our extension increases the benchmark to 381 entries.[2] The extended benchmark covers 172 unique lemmas. The sense inventory is distributed as follows: 145 word types have two senses, 22 have three senses, 4 have four senses, and 1 word type has seven senses.

The benchmark can be accessed for research purposes under an open license, with images licensed under Creative Commons. We release the benchmark annotations (queries, trigger phrases, candidate lists, and gold labels) under an open research license at [3].

This extension enables repeatable benchmarking of multilingual CLIP checkpoints and multimodal LMMs on Ukrainian Visual-WSD, and supports controlled ablations in later sections.

Benchmark format. We follow the format of SemEval-2023 Visual-WSD (Raganato et al., 2023). Each instance consists of: (i) an ambiguous target word and a short trigger phrase that specifies the intended sense, and (ii) a set of ten candidate images containing one gold image depicting the intended sense and nine distractors. Distractors are sampled from three categories: (1) images depicting alternative senses of the same word type, (2) images depicting semantically related concepts, and (3) unrelated random concepts. This candidate construction encourages sense-level grounding rather than generic semantic matching.

Details of the annotation protocol, validation, and annotator effort are provided in Appendix A.

3.1 Prediction

Given a query phrase and ten candidate images, embedding-based models rank candidates by cosine similarity in the shared image-text embedding space (e.g., CLIP-style retrieval), and the top-ranked image is selected as the prediction. When direct access to embeddings is unavailable (e.g., GPT-5.1), we prompt the model to rank the ten candidates by semantic relevance to the query and use the top-ranked image as the prediction; details are provided in Section 3.4.

3.2 Evaluation metrics

We report two standard retrieval metrics: HIT@1 and Mean Reciprocal Rank (MRR). HIT@1 measures the percentage of instances where the gold image is ranked first. MRR averages the inverse rank of the gold image across instances, giving partial credit when the gold image is ranked near the top.

3.3 Multilingual CLIP benchmarking

We benchmark several multilingual CLIP models on the extended Ukrainian Visual-WSD benchmark, comparing their performance with English, Farsi, and Italian baselines from the SemEval-2023 Visual-WSD task.

Table 1 shows that off-the-shelf multilingual CLIP checkpoints fall short on Ukrainian Visual-WSD: the best model achieves 43.82 HIT@1 / 60.6 MRR. We additionally report the SemEval-2023 English CLIP baseline to contextualize the task and provide an upper reference for zero-shot CLIP performance in the original benchmark setting. This reference is used to illustrate the potential for further improvement, not to draw cross-language conclusions.

[2] Each entry corresponds to one ambiguous word type in a specific sense, paired with a short sense-disambiguating trigger phrase and a fixed set of ten candidate images.

[3] https://huggingface.co/datasets/yuriilaba/ukrainian-visual-wsd-benchmark

Model	Image+text Params (M)	HIT@1	MRR
LAION CLIP ViT-H/14 XLM-R large	1193.0	**43.82**	**60.6**
M-CLIP ViT-B/16+ XLM-R large	677.8	37.80	56.7
LAION CLIP ViT-B/32 XLM-R base	366.1	37.00	54.4
M-CLIP ViT-L/14 XLM-R large	864.6	36.48	55.5
M-CLIP ViT-L/14 LaBSE	775.5	35.17	53.96
M-CLIP ViT-B/32 XLM-R large	648.3	34.65	53.9
Sentence-Transformers CLIP ViT-B/32	135.1	28.87	48.74
English baseline†	135.1	60.48	73.88
Farsi baseline†	135.1	28.50	46.70
Italian baseline†	135.1	22.62	42.61

Table 1: HIT@1 and MRR metrics for various multilingual CLIP models on the extended Ukrainian Visual-WSD benchmark. Baseline results for English, Farsi, and Italian from SemEval-2023 Visual-WSD, obtained with the the CLIP ViT-B/32† model, are also included for comparison.

3.4 LLM benchmarking

To extend benchmarking beyond embedding-based retrieval models, we evaluate a set of LMMs from multiple model families (OpenAI, Google Gemini, Alibaba Qwen, and Anthropic) on the extended Ukrainian Visual-WSD benchmark.

We prompt models with the ambiguous word, trigger phrase, and ten images, and ask for a ranked list of image indices (Appendix B). We treat the top-ranked image as the prediction and report HIT@1 and MRR.

Full LLMs benchmarking results are reported in Table 2. We used a fixed temperature of 0.01 for all LLMs evaluations and keep decoding settings constant across models.

Family	Model	HIT@1	MRR	Image Detail
OpenAI	GPT-4o	**67.6**	**79.7**	LOW
	GPT-4o	56.0	71.0	HIGH
	GPT-4o	56.5	72.9	AUTO
	GPT-4.1	**71.4**	**81.9**	LOW
	GPT-4.1	71.0	81.1	HIGH
	GPT-4.1	71.0	81.1	AUTO
	GPT-4.1-mini	40.4	58.6	LOW
	GPT-4.1-mini	**42.7**	**59.5**	HIGH
	GPT-4.1-mini	41.9	58.7	AUTO
	GPT-4.1-nano	19.7	38.4	LOW
	GPT-4.1-nano	**19.9**	**39.0**	HIGH
	GPT-4.1-nano	19.3	37.8	AUTO
	gpt-5.2	**63.4**	**74.5**	LOW
	gpt-5.2	-	-	HIGH
	gpt-5.2	62.5	74.5	AUTO
	gpt-5.2-mini	**67.1**	**78.6**	LOW
	gpt-5.2-mini	-	-	HIGH
	gpt-5.2-mini	-	-	AUTO
	gpt-5.2-nano	37,1	53	LOW
	gpt-5.2-nano	-	-	HIGH
	gpt-5.2-nano	**37.8**	**53.9**	AUTO
Gemini	Gemini-1.5-pro	**60.9**	**73.9**	LOW
	Gemini-1.5-pro	59.3	73.1	HIGH
	Gemini-1.5-pro	–	–	AUTO
Qwen	Qwen2.5-VL-7B	**38.6**	**55.7**	LOW
	Qwen2.5-VL-7B	34.6	52.4	HIGH
	Qwen2.5-VL-7B	38.1	54.6	AUTO
	Qwen2.5-VL-32B	41.0	75.4	LOW
	Qwen2.5-VL-32B	–	–	HIGH
	Qwen2.5-VL-32B	–	–	AUTO
Anthropic	Claude 3.5 Sonnet	63.0	75.5	LOW
	Claude 3.5 Sonnet	60.4	74.0	HIGH
	Claude 3.5 Sonnet	–	–	AUTO

Table 2: Performance of LLM families on the extended Ukrainian Visual-WSD benchmark.

4 Training Data and Dataset Construction

Our next goal is to study data-centric, compute-efficient adaptation of CLIP-style models for Ukrainian Visual-WSD. To this end, we fine-tune the same backbone under a fixed training budget constant (epochs/steps) and vary the source of image–text dataset: from broad, generic image-text pairs to sense-aligned Wikipedia-derived supervision designed to explicitly target disambiguation. This section describes the training corpora used in our experiments and the two dataset construction pipelines we propose for mining sense-grounded Ukrainian supervision from Wikipedia. Table 3 lists the four training corpora used in our experiments.

4.1 Sense-aligned Wikipedia (UA) construction (SenseWiki-UA)

The absence of dedicated training data for Ukrainian Visual-WSD presents a fundamental challenge for model adaptation. To address this limitation, we construct SenseWiki-UA, a sense-aligned image–text dataset that links dictionary senses of ambiguous words to sense-specific pages in Ukrainian Wikipedia. By aligning dictionary and Wikipedia definitions, we harvest images explicitly grounded in a particular meaning, providing more

Dataset	#Pairs
SenseWiki-UA	27,543
RA-Wiki-UA	38,100
LAION-UA	62,649
Multi30K-UA	28,905

Table 3: Training corpora used for contrastive fine-tuning. SenseWiki-UA and RA-Wiki-UA are Wikipedia-derived and sense-aligned; LAION-UA and Multi30K-UA provide generic Ukrainian image-text supervision.

Quantity	Value
Dictionary headwords	60,620
Dictionary Homonyms ($\geq$ 2 meanings)	1,254
Dictionary Homonyms senses	2,638
Headwords covered by Wikipedia	7,557
Homonyms covered by Wikipedia	301
Homonym senses covered by Wikipedia	408
SenseWiki-UA image-text pairs	27,543

Table 4: SenseWiki-UA scale and Wikipedia coverage.

targeted supervision than generic web captions.

We start from a Ukrainian dictionary (ULIF-NASU, 2010) with 60,620 headwords. Among them, 1,254 are homonyms (words with $\geq$ 2 meanings), which together account for 2,638 distinct sense entries. As a complementary source of grounded multimodal evidence, we use the Ukrainian Wikipedia dump (pages, lead definitions/summaries, images, and captions when available).

Overview. SenseWiki-UA is built in four stages: (1) sense inventory traversal from the dictionary, (2) Wikipedia candidate retrieval for each headword, (3) definition alignment to select the best-matching Wikipedia sense, and (4) image harvesting with caption fallback and filtering.

Stage 1: Dictionary sense inventory. We iterate over all dictionary headwords (including non-homonyms) and their sense definitions. Although the primary target is homonym disambiguation, traversing the full dictionary provides a uniform pipeline and allows Wikipedia alignment to fail gracefully when ambiguity is absent or Wikipedia coverage is missing.

Stage 2: Wikipedia candidate retrieval. For each headword, we retrieve candidate Wikipedia pages/sections corresponding to that word (including disambiguation-style pages and sense-specific entries when present). From each candidate, we extract a short definition (e.g., lead sentence/summary) and associated multimedia metadata (images and captions, if provided).

Stage 3: Definition alignment (sense linking). To link a dictionary sense to Wikipedia, we compute cosine similarity between the dictionary definition and each candidate Wikipedia definition in a shared embedding space. We select the closest Wikipedia definition as the aligned sense. To control noise, we discard alignments with similarity < 0.5 [4].

Stage 4: Image harvesting. For each accepted alignment, we follow the aligned Wikipedia page and harvest its images. If an image has a Wikipedia caption, we use it as supervision; otherwise, we fall back to the dictionary definition for the aligned sense as the textual description. This produces sense-conditioned image-text pairs without manual annotation.

SenseWiki-UA initially yields 31,699 candidate image–text pairs spanning 11,629 dictionary headwords. Due to limited Ukrainian Wikipedia coverage, sense alignment is possible for only a subset of the dictionary: of 1,254 homonyms (2,638 senses), we align 405 homonyms (596 senses) to Wikipedia sense pages.

Quality filtering further reduces the dataset. 13% of the original 31,699 entries link to Wikipedia pages with no usable images and are removed. Additionally, 20% of the original 31,699 entries come from pages whose images lack captions; for these, we use the aligned dictionary definition as the text. After filtering, SenseWiki-UA contains 27,543 image–text pairs covering 7,557 headwords, 301 homonyms, and 408 homonym senses (Table 4).

SenseWiki-UA addresses two requirements for Visual-WSD training: (i) it anchors supervision to a dictionary-defined sense via definition alignment, and (ii) it grounds that sense in real images from Ukrainian Wikipedia. Its scale is bounded by Wikipedia's illustration and caption coverage.

Manual quality audit. To estimate this noise, we manually audited 100 randomly sampled image-text pairs after filtering (Table 5). For each pair, we assessed dictionary-Wikipedia sense alignment, image grounding, text quality, and overall usability

[4]We set the similarity threshold to 0.5 based on a manual audit of 200 randomly sampled alignments across the score range, where scores below 0.5 showed frequent sense mismatches.

Dimension	Label	%
Sense alignment	Correct	88.0
	Wrong	12.0
Image grounding	Clear visual grounding	82.0
	Weak/indirect grounding	2.0
	Unrelated/misleading	14.0
	Missing/broken image	2.0
Text quality	Good visual text	50.0
	Generic/definition-like	22.0
	Caption too short	6.0
	Wrong/misleading	22.0
Overall quality	Usable	68.0
	Noisy but usable	12.0
	Not usable	20.0

Table 5: Manual audit of 100 randomly sampled SenseWiki-UA image–text pairs.

for contrastive training. The audit confirms that most pairs are suitable for training, with residual noise mainly due to imperfect sense linking, weak visual grounding, and low-specificity textual supervision.

4.2 Retrieval-augmented Wikipedia (UA) construction (RA-Wiki-UA)

RA-Wiki-UA, a semi-supervised augmentation dataset that leverages (i) the structure of our target benchmark and (ii) external multimodal evidence from Ukrainian Wikipedia. The key idea is to reduce domain shift by sampling training images that are close to benchmark instances in a pretrained vision–language embedding space, thereby concentrating supervision on visually relevant concepts.

Overview. The pipeline has three stages: (1) harvesting multimodal metadata from Ukrainian Wikipedia, (2) similarity-based retrieval to form a visually relevant candidate pool, and (3) LLM-based caption generation to produce diverse Ukrainian supervision.

Stage 1: Wikipedia harvesting. We collect 150,000 images from the Ukrainian Wikipedia dump. Only 64,130 include Ukrainian captions; for all images, we also extract the page title and lead summary. This sparsity is a key limitation we also encountered in SenseWiki-UA: many Ukrainian pages are weakly illustrated and image captions are often missing. Because Wikipedia is a common backbone for multimodal pretraining and adaptation, this missing metadata substantially limits the amount of reliable Ukrainian supervision we can extract without augmentation.

Component	#Images	#Texts
Ukrainian Visual-WSD benchmark	381	381
Wikipedia harvest	150,000	64,130
RA-Wiki-UA (synthetic pairs)	3,810	38,100

Table 6: Scale of components used to construct RA-Wiki-UA.

Stage 2: Similarity-driven candidate selection. For each benchmark gold image I_t (one per Visual-WSD instance), we retrieve the top-10 most similar Wikipedia images using cosine similarity in the embedding space of OpenCLIP ViT-H/14 + XLM-R large (LAION-5B).[5] Across the 381 benchmark instances, this retrieval produces 3,810 Wikipedia images that are visually close to benchmark sense depictions.

Stage 3: Sense-aware caption generation. To obtain training pairs for contrastive tuning, we generate Ukrainian captions for each retrieved image by prompting GPT-5.1[6] with multimodal inputs: the image itself and its associated Wikipedia context (page title, available caption, and article summary). The prompt: (i) identifies ambiguous concepts in the image, (ii) disambiguates them using the visual–text context, and (iii) generates multiple diverse Ukrainian captions. The full prompt is provided in Appendix C.

In total, RA-Wiki-UA contains 38,100 synthetic image-text training pairs. Table 6 summarizes the dataset sizes across RA-Wiki-UA construction.

We assess retrieval quality by measuring cosine similarity between each benchmark image and its retrieved Wikipedia neighbors in the retrieval embedding space. For most instances, even the tail of the top-10 remains close: the median similarity is 0.686 for the nearest neighbor and 0.572 for the 10th neighbor (Table 7).

We also quantify within-image caption diversity using Distinct-n and Self-BLEU, computed over the set of captions generated for the same image. Averaged across 381 images, Distinct-1/2 reach 0.46/0.85, while Self-BLEU is 0.038 (Table 8), showing that the ten captions per image are typically quite different from one another.

[5]We use this checkpoint as the retrieval backbone due to its strongest zero-shot performance among multilingual CLIP variants on our Ukrainian Visual-WSD benchmark (Table 1).

[6]We use GPT-5.1 because our university grant provided funding for OpenAI API usage.

Statistic	s_1	s_{10}	$s_1 - s_{10}$
Mean	0.684	0.565	0.120
Median	0.686	0.572	0.116
P10	0.578	0.470	0.025
P90	0.804	0.642	0.209

Table 7: Retrieval quality for RA-Wiki-UA. s_1 and s_{10} denote cosine similarity to the top-1 and top-10 retrieved Wikipedia images for each benchmark instance in the retrieval embedding space.

Metric	Mean	Median
Distinct-1 (within image)	0.458	0.459
Distinct-2 (within image)	0.846	0.853
Self-BLEU (within image)	0.038	0.035
Distinct-1 (overall)	0.103	
Distinct-2 (overall)	0.611	

Table 8: Diversity of synthetic Ukrainian captions in RA-Wiki-UA. Within-image metrics are computed over captions generated for the same image and then averaged across images.

RA-Wiki-UA addresses two practical constraints in Ukrainian Visual-WSD adaptation. First, nearest-neighbor retrieval around benchmark images yields training visuals that match the benchmark distribution without manual labeling. Second, context-conditioned Ukrainian captions (10 per image) provide richer lexical and syntactic variation than Wikipedia metadata, which is often missing or templatic.

RA-Wiki-UA is a benchmark-targeted supervision source, not a fully benchmark-independent pretraining corpus. Although benchmark images are not reused for training, retrieval is initialized from the visual space of benchmark gold images. This lets us test whether sense-aligned Wikipedia images improve Ukrainian Visual-WSD, but it also limits the conclusions: some gains may reflect benchmark-specific visual alignment. We therefore interpret RA-Wiki-UA as evidence for sense-grounded supervision in a controlled adaptation setting and leave benchmark-independent retrieval for future work.

Manual quality audit. Because RA-Wiki-UA combines similarity-based image retrieval with generated Ukrainian sentences, we manually audited 100 randomly sampled image-sentence pairs after construction (Table 9). The audit shows that 72.0% of sampled pairs were usable for contrastive training, while 28.0% were not usable. The main failure mode was retrieval drift: visually mismatched retrieved images led to wrong or unclear senses and misleading generated sentences.

Dimension	Label	%
Image relevance	Correct image	72.0
	Wrong image	28.0
Sense correctness	Correct sense	72.0
	Wrong/unclear sense	28.0
Sentence quality	Good sentence	36.0
	Generic but usable	36.0
	Wrong/misleading	28.0
Overall quality	Usable	72.0
	Not usable	28.0

Table 9: Manual audit of 100 randomly sampled RA-Wiki-UA image–sentence pairs.

4.3 External Ukrainian corpora

LAION-UA We construct LAION-UA by filtering LAION-2B-multi, the multilingual LAION-2B subset of the LAION dataset family (Schuhmann et al., 2022), to Ukrainian image-text pairs. Since LAION-2B-multi is highly imbalanced across languages, continued tuning on the Ukrainian slice upweights Ukrainian captions and shifts the model toward the target-language distribution (Gururangan et al., 2020). We use LAION-UA as a generic, non-sense-aligned baseline from the same data source family as pre-training, isolating gains from additional Ukrainian web data versus sense-aligned supervision (SenseWiki-UA, RA-Wiki-UA).

We start from laion/laion2B-multi, which provides metadata and URLs for image-text pairs.[7] We then apply the following filters: (i) retain only entries with LANGUAGE = uk; (ii) stream the first 20M records from this Ukrainian subset; and (iii) keep only pairs whose provided similarity score exceeds 0.33. [8] This yields 62,649 Ukrainian image-text pairs, which we use as a generic, non-sense-aligned Ukrainian training corpus.

Multi30K-UA. We also use the Ukrainian extension of Multi30K (Saichyshyna et al., 2023), a curated set of image descriptions with manually verified Ukrainian translations. We include it as a clean, high-precision contrastive tuning corpus that complements web-scale supervision. Following standard preprocessing, we drop a small number of

[7]The dataset is distributed as a table of metadata with image URLs rather than packaged images.

[8]We use LAION's released similarity score as a lightweight quality filter, keeping only pairs above 0.33 as a conservative cutoff to exclude weak or noisy image-text matches.

entries with missing/corrupted text fields, leaving 28,905 image-text pairs.

5 Experimental setup

For subsequent adaptation experiments, we use laion/CLIP-ViT-B-32-xlm-roberta-base: it is much smaller (366M vs. 1.19B for ViT-H/14 + XLM-R large) while remaining competitive in zero-shot (Table 1), making it a practical backbone for data and freezing ablations.

We run four single-corpus training runs: LAION-UA, Multi30K-UA, SenseWiki-UA, and RA-Wiki-UA. We also train a size-matched mixture: we sample 26,165 pairs per corpus (min corpus size) and concatenate to control for distribution:

- SenseWiki-UA + LAION-UA + RA-Wiki-UA + Multi30K-UA: combines sense-grounded and generic supervision to test whether mixing sources improves Visual-WSD.

Finally, we train two full-corpus mixtures (no subsampling) as controls:

- SenseWiki-UA + RA-Wiki-UA: to measure how far targeted supervision goes without generic web data.
- LAION-UA + Multi30K-UA: to separate gains from data volume/caption diversity from explicit sense alignment.

For compute-efficient adaptation, we freeze the vision tower and tune only the top k text Transformer layers, sweeping k from 0 to 12 to quantify the accuracy-capacity trade-off under identical training conditions; k=0 tunes only the text and image projection layers.

Ukrainian Visual-WSD serves as our primary evaluation target. We do not train on benchmark instances; instead, we evaluate once per epoch during fine-tuning and select the checkpoint with the highest Ukrainian Visual-WSD performance, reporting HIT@1 and MRR for that checkpoint. To confirm that improved Visual-WSD performance does not reduce overall image–text alignment, we additionally measure retrieval robustness on held-out 5% splits of the training corpora (LAION-UA and Multi30K-UA) using mean recall, defined as the average of R@1/5/10 over both image-to-text and text-to-image retrieval directions, where R@K is the fraction of queries whose true match appears in the top-K retrieved results.

Table 12 summarizes the fine-tuning configuration used across experiments.

6 Results

Visual-WSD results. Across training setups, most of the improvement is driven by sense-grounded supervision. Generic data is mixed: Multi30K-UA helps slightly, while LAION-UA can hurt, and generic-only mixtures stay close to baseline (See Table 10).

SenseWiki-UA increases HIT@1 by 2.11% with only 6.14% trainable parameters (best $k = 3$), indicating that sense anchoring alone provides a useful supervision signal for disambiguation. RA-Wiki-UA improves further (+4.47% HIT@1), but it requires a much larger update budget (best $k = 12$, 23.56% trainable).

The best result comes from SenseWiki-UA + RA-Wiki-UA (best $k = 12$, 23.56% trainable), delivering a +6.05% HIT@1 improvement and showing that the two pipelines are complementary rather than redundant. By contrast, LAION-UA slightly underperforms the baseline (-0.23 HIT@1), and the generic-only mixture (LAION-UA + Multi30K-UA) stays near baseline (+0.80 (%) HIT@1). Overall, this highlights a limitation of generic supervision: adding more Ukrainian captions alone is not enough: without explicit sense structure, it does not reliably improve sense discrimination.

We provide a breakdown by sense ambiguity comparing the baseline to our best-performing fine-tuned setup, along with a paired significance test, in Appendix E.

Freezing ablation. Varying k (vision frozen; tune the top-k text blocks) reveals that each dataset has a different best-performing k under the accuracy–capacity trade-off (Figure 1).

SenseWiki-UA peaks early ($k = 3$) and then saturates, while RA-Wiki-UA and SenseWiki-UA + RA-Wiki-UA keep improving up to $k = 12$, suggesting that richer caption supervision benefits from deeper text-tower adaptation. Generic corpora are less stable: Multi30K-UA peaks later ($k = 9$) with limited improvement, whereas LAION-UA remains below baseline. Overall, the ablation quantifies the parameter-efficiency trade-off: limited tuning can help (6.14% trainable at $k = 3$), but the best Visual-WSD results require deeper updates (23.56% at $k = 12$).

Trade-off: Visual WSD accuracy vs retrieval robustness. The accuracy-robustness comparison (Figure 2) shows that the strongest Visual-WSD setup slightly reduces retrieval robustness, reflect-

Training data	Best K	Trainable params (%)	# train samples	HIT@1(%)	MRR(%)
Baseline model (pretrained)	-	-	-	37.00	54.29
SenseWiki-UA	3	6,14	27,543	39.11	55.70
RA-Wiki-UA	12	23,56	38,100	41.47	57.43
LAION-UA	8	15,82	59,516	36.77	54.36
Multi30K-UA	9	17,75	27,941	38.58	55.72
SenseWiki-UA + RA-Wiki-UA	12	23,56	64,326	**43.05**	**58.26**
SenseWiki-UA + LAION-UA + RA-Wiki-UA + Multi30K-UA	8	15.82	104,660	40.68	56.63
LAION-UA + Multi30K-UA	1	2,27	87,457	37.80	55.05

Table 10: Ukrainian Visual-WSD results (HIT@1, MRR). Best K is the number of top text-encoder Transformer layers tuned (vision encoder frozen) that yields the highest Visual-WSD score for each training setup.

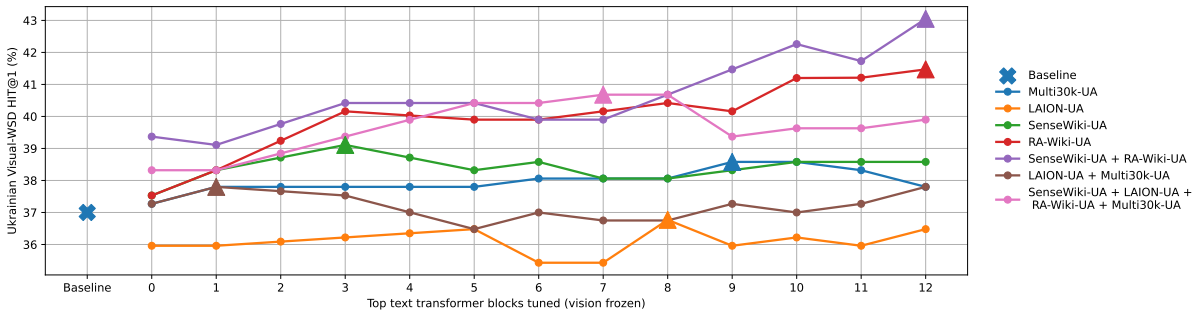

Figure 1: Freezing ablation across training datasets.

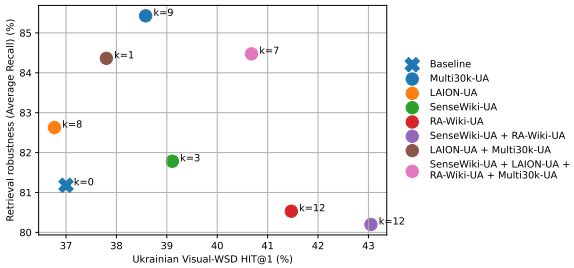

Figure 2: Accuracy-robustness trade-off at the best tuning setting.

ing a small trade-off in overall image–text alignment when optimizing for Visual-WSD. Generic supervision shows the opposite trend: it preserves or even improves robustness, but yields much smaller WSD gains. The full mixture sits between these extremes, suggesting that under a fixed training budget, adding generic data can improve robustness while weakening the sense-grounded training signal that drives WSD performance.

7 Conclusions

We show that improving Ukrainian Visual-WSD is primarily a data alignment problem: adding more Ukrainian image-text pairs is not sufficient unless the supervision is explicitly grounded in sense structure. To enable reliable evaluation in this low-resource setting, we extend the Ukrainian Visual-WSD benchmark from 87 to 381 instances and establish reference results for multilingual CLIP checkpoints and multimodal LMMs.

To address the supervision bottleneck, we introduce two scalable Wikipedia-derived pipelines: SenseWiki-UA (dictionary-to-Wikipedia sense linking) and RA-Wiki-UA (retrieval-neighborhood Wikipedia images with generated Ukrainian captions). With compute-efficient tuning (vision frozen; tune top-k text layers), combining them improves HIT@1 from 37.00% to 43.05% (+6.05), while generic-only Ukrainian corpora remains near the baseline.

Although this still trails LMM performance, it provides a lightweight alternative for deployment: our best model retains retrieval-style inference and runs orders of magnitude faster than API-based LMM ranking (Appendix F).

To support reproducibility, we release dataset construction scripts, and fine-tuning code.[9]

[9] https://github.com/YuriiLaba/ukrainian-visual-wsd

Limitations

Our study is scoped to Ukrainian. While both Wikipedia-based pipelines are largely language-agnostic in design, transferring them to other low-resource languages would require (i) a compatible sense inventory, (ii) sufficient Wikipedia coverage and illustration density. Performance improvements and conclusions may not carry over under weaker resources.

Although we extend Ukrainian Visual-WSD to 381 instances, it still covers a limited set of lemmas and senses, with a skew toward 2-sense homonyms.

RA-Wiki-UA uses GPT-5.1 to generate Ukrainian captions, so dataset quality depends on model behavior, prompt design, and the provided Wikipedia context.

References

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Yaping Chai, Haoran Xie, and Joe S Qin. 2025. Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. *arXiv preprint arXiv:2501.18845*.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023. AltCLIP: Altering the language encoder in CLIP for extended language capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8666–8682, Toronto, Canada. 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.

Krishno Dey, Prerona Tarannum, Md Arid Hasan, Imran Razzak, and Usman Naseem. 2024. Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings. *arXiv preprint arXiv:2410.13153*.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Hyunjae Kim, Seunghyun Yoon, Trung Bui, Handong Zhao, Quan Tran, Franck Dernoncourt, and Jaewoo Kang. 2024. Fine-tuning CLIP text encoders with two-step paraphrasing. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2175–2184, St. Julian's, Malta. Association for Computational Linguistics.

Teerath Kumar, Rob Brennan, Alessandra Mileo, and Malika Bendechache. 2024. Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access*.

Yurii Laba, Yaryna Mohytych, Ivanna Rohulia, Halyna Kyryleyza, Hanna Dydyk-Meush, Oles Dobosevych, and Rostyslav Hryniv. 2024. Ukrainian visual word sense disambiguation benchmark. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 61–66, Torino, Italia. ELRA and ICCL.

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

Zheng Li, Lijia Si, Caili Guo, Yang Yang, and Qiushi Cao. 2024. Data augmentation for text-based person retrieval using large language models. *arXiv preprint arXiv:2405.11971*.

Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavalle, and Roberto Navigli. 2025. Do large language models understand word senses? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33885–33904.

Antonina Mijatovic, Davide Buscaldi, and Ekaterina Borisova. 2023. RCLN at SemEval-2023 task 1: Leveraging stable diffusion and image captions for visual WSD. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2174–2178, Toronto, Canada. Association for Computational Linguistics.

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.

Michael Ogezi, Bradley Hauer, Talgat Omarov, Ning Shi, and Grzegorz Kondrak. 2023. UAlberta at SemEval-2023 task 1: Context augmentation and translation for multilingual visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2043–2051, Toronto, Canada. Association for Computational Linguistics.

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

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.

Nataliia Romanyshyn, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2024. Automated extraction of hypo-hypernym relations for the Ukrainian WordNet. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 51–60, Torino, Italia. ELRA and ICCL.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

ULIF-NASU. 2010. Словник української мови [Dictionary of the Ukrainian language], volume 20 of Словники України [Dictionaries of Ukraine]. Наук. думка [Nauk. dumka], Kyiv.

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.

Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. 2022. Image data augmentation for deep learning: A survey. arXiv preprint arXiv:2204.08610.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18123–18133.

Yue Zheng, Yuhao Chen, Bin Qian, Xiufang Shi, Yuanchao Shu, and Jiming Chen. 2025. A review on edge large language models: Design, execution, and applications. ACM Computing Surveys, 57(8):1–35.

Yao Zhu, Yuefeng Chen, Xiaofeng Mao, Xiu Yan, Yue Wang, Wang Lu, Jindong Wang, and Xiangyang Ji. 2024. Enhancing few-shot clip with semantic-aware fine-tuning. IEEE Transactions on Neural Networks and Learning Systems.

A Annotation protocol, quality control, and human effort

A.1 Annotation protocol and quality control

All word senses and image associations were curated by a team of three professional Ukrainian linguists under the supervision of a lead annotator. We followed the same core annotation principles and labeling protocol as in Laba et al. (2024).

Guidelines. For each word type and sense, annotators:

- selected/verified the target sense using the Ukrainian homonym dictionary definition;
- wrote a short trigger phrase that disambiguates the intended sense while keeping context minimal (SemEval-style);
- selected a gold image that clearly depicts the intended sense (unambiguous depiction, central concept visible);
- selected distractors from three categories: other senses of the same word, semantically

Statistic	Value
Instances	381
Flagged & adjudicated	29
Revised / replaced	16
Agreement (P_o)	92.4%
Cohen's κ	0.7

Table 11: Validation outcomes and agreement from the binary validation decision (ACCEPT vs. NEEDS-REVIEW).

related concepts, and unrelated random concepts, ensuring distractors do not depict the target sense;

- avoided near-duplicates within the same candidate set (e.g., the same scene/object with minor crops);
- verified that the gold image and distractors are culturally and linguistically appropriate for Ukrainian queries.

Annotation quality and agreement. Dataset creation followed a two-stage workflow: (i) initial construction by one annotator and (ii) validation by a second annotator, with lead-annotator adjudication for flagged cases. Validation used a binary decision (ACCEPT vs. NEEDS-REVIEW); Table 11 summarizes outcomes and agreement.

A.2 Human effort and compensation

The benchmark extension required substantial expert annotation effort. In total, annotation and validation took approximately 127 person-hours across three linguists over six weeks. On average, annotators spent about 20 minutes per instance (including trigger phrasing, candidate selection, and validation). Annotators were compensated through a dedicated university grant under standard institutional procedures.

A.3 Human data collection and ethics

Annotators were recruited as professional Ukrainian linguists via Ukrainian Catholic University, based on expertise in Ukrainian lexicography. Compensation was provided through a dedicated university grant under standard institutional procedures.

Annotators were compensated at $12/hour, and the lead annotator at $18/hour.

All annotators provided informed consent to participate in paid annotation, including consent for their annotations (sense labels, trigger phrases, and image candidate sets) to be used for research and released in anonymized form; no sensitive personal, behavioral, or biometric data were collected.

Our institution determined that this work does not constitute human-subjects research because it involves paid expert annotation with no collection of personal or sensitive data beyond professional role; therefore IRB review was not required.

B LLM Visual-WSD prompt template

```
[Instruction]
You are an expert in semantic analysis
    of polysemous Ukrainian words.
Given an ambiguous word, a short trigger
     phrase, and 10 candidate images:

(1) Infer the intended sense of the
    ambiguous word from the trigger
    phrase.
(2) Rank the 10 images by how well they
    depict that sense.

[Output format]
Return ONLY valid JSON (no extra text):
{"ranked_images":[i1,i2,i3,i4,i5,i6,i7,
    i8,i9,i10]}
where i1..i10 are a permutation of
    integers 1..10 (no repeats).

[Input]
word: "{WORD}"
trigger: "{TRIGGER}"
images: 1..10
```

C Prompt For RA-Wiki-UA Sentences Generation

```
[Task] Multimodal Sentence Generation
    for Visual Word Sense Disambiguation
     (WSD)
[Language] Ukrainian

[Input modalities]
- Page Title: {page_title}
- Image Caption: {image_caption or "N/A
    "}
- Page Summary: {page_summary or "N/A"}

[Objective]
Generate diverse and natural Ukrainian
    sentences that express the specific
    sense of a target word visually
    depicted in an image. The target
    word may be found in any of the
    three sources (title, caption, or
    summary) and should be disambiguated
     from its other meanings using the
    image and text.

[Step-by-step instructions]

1. Analyze the image
- Identify salient visual elements (
    objects, actions, environment,
    attributes).
```

```
- Describe what is clearly shown, even
    if not mentioned in the text.

2. Identify the target word
- Scan the title, caption, and summary
    to find a word that:
  - Has multiple meanings
  - Is visually represented in the image

3. Disambiguate the sense
- Use the caption to understand the
    intent behind the image.
- Use the summary to identify other
    possible meanings.
- Choose the sense that best matches the
     visual content.

4. Generate 10 sentences
- All sentences must:
  - Refer to the visual sense of the
      target word
  - Include either the target word or a
      valid synonym
  - Contain visual context grounded in
      the image
- Vary syntax and collocations:
  - Use adjectives, adverbs,
      prepositional phrases
  - Shift grammatical roles (subject,
      object, etc.)
  - Include passive and active voice
- At least 5 of the 10 sentences must
    use synonyms, not the target word
    itself.

5. Select synonyms thoughtfully
- Use only synonyms that express the
    exact same visual sense.
- Acceptable types:
  - Direct lexical synonyms
  - Semantic equivalents or paraphrases
  - Hypernyms only if natural and
      necessary
- Avoid:
  - Synonyms tied to other senses
  - Overly abstract or obscure
      alternatives

6. Optional contrastive examples
- Provide 1-2 sentences using the same
    word but showing a different sense
    from the image.
- These help highlight disambiguation
    clearly.


[Output format]
Return ONLY valid JSON (no extra text):
{
  "target_word": "...",
  "visual_sense_description": "...",
  "key_visual_elements": ["..."],
  "sentences": [
    "..."
  ],
  "synonyms_used": ["..."],
  "contrastive_sentences": [
    "..."
  ]
}

[Final checklist]
1) Does the target word have multiple
    senses?
2) Is the visual sense clearly matched
    and described?
3) Are at least 5 sentences using proper
     synonyms?
4) Are visual elements embedded in every
     sentence?
5) Do the sentences sound natural in
    Ukrainian?
```

D Fine-Tuning Parameters

Parameter	Value
Backbone	xlm-roberta-base-ViT-B-32
Pretrained weights	laion5b_s13b_b90k
Text tower (XLM-R base)	12 blocks; train top-k blocks
Vision tower (ViT-B/32)	12 blocks; frozen in all adaptation runs
Training objective	Symmetric contrastive loss
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	0.1
Epochs	5
Batch size	512
Gradient clipping	max norm = 1.0
LR schedule	CosineAnnealingLR
Random seed	42

Table 12: Fine-tuning configuration used in our experiments.

E Additional Performance Analysis

We complement the main results with analyses that probe when fine-tuning helps most and whether the observed gains are statistically reliable.

E.1 Performance by sense ambiguity

Figure 3 analyzes how Ukrainian Visual-WSD performance varies with lexical ambiguity, measured as the number of senses for the target lemma. We compare the frozen baseline against our best fine-tuned model and report both MRR (left) and HIT@1 (right). Fine-tuning improves retrieval across all ambiguity levels, with the largest gains observed for highly ambiguous lemmas (4 and 7 senses).

E.2 Statistical Significance of the Obtained Results

To test whether the improvement is systematic rather than driven by a small number of instances, we apply a one-sided Wilcoxon signed-rank test to per-instance reciprocal rank on the 381 evaluation instances shared by both models. The alternative hypothesis is that the fine-tuned model yields higher reciprocal rank than the frozen baseline. The test returns W=9409.0 with p=1.25 $\times$ 10^{-4},

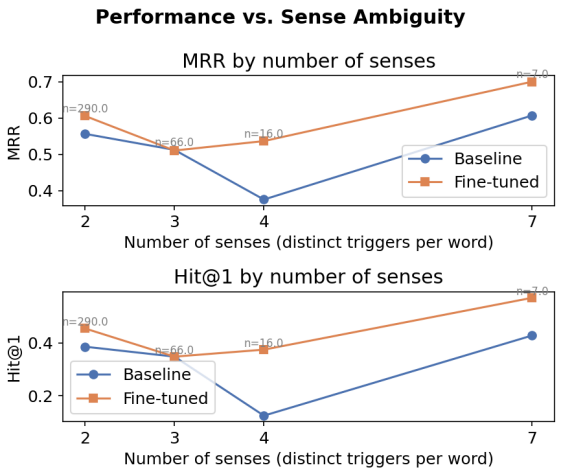

Figure 3: Performance vs. Sense Ambiguity.

indicating a statistically significant improvement at α=0.05. We use Wilcoxon because it is non-parametric and does not assume a specific score distribution, which is appropriate for bounded, non-Gaussian retrieval outcomes such as reciprocal rank.

F Inference Latency and API Cost

Model	HIT@1	Latency (ms)	API Price ($)
Gemini-1.5-pro	73.9	23432.67	0.004
GPT-4.1	71.4	24684.74	0.003
GPT-4o	67.6	22696.09	0.004
Ours (best fine-tuned)	**43.05**	**8.27**	**–**
GPT-4.1-mini	42.7	21831.47	0.003
GPT-4.1-nano	19.9	19012.27	0.003

Table 13: Performance and efficiency comparison of LLMs and our best fine-tuned CLIP-based model on the Visual-WSD benchmark. All metrics are reported per average retrieval from the V-WSD benchmark.