From Benchmark to Better Embeddings: Leveraging Synonym Substitution to Enhance Multimodal Models in Ukrainian

Volodymyr Mudryi*
Ukrainian Catholic University
Lviv, Ukraine
mudryi.pn@ucu.edu.ua

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

Abstract

We study the robustness of text–image retrieval for Ukrainian under synonym-substitution attacks (SSA). On Multi30K with OpenCLIP, we evaluate two SSA methods: dictionary-based and LLM-based, and find Ukrainian degrades far more than English (e.g., GPT-4o SSA drops HIT@1 from 32.1 → 10.9 vs. 41.6 → 30.4). We introduce a Hybrid method that filters dictionary candidates with an LLM to preserve sense and grammar, yielding higher-quality perturbations (Ukrainian HIT@1 16.8 vs. 7.6/10.9). To mitigate this problem, we propose synonym-augmented fine-tuning, injecting one-word substitutions into training; it boosts robustness (Hybrid 28.1, GPT-4o 25.1) without harming original performance. This is the first systematic SSA evaluation for Ukrainian multimodal retrieval and a practical recipe for improving models in low-resource, morphologically rich languages. We release code, prompts, and trained checkpoints at github.com/YuriiLaba/UA-B2BE.

1 Introduction

Embeddings provide vector-space representations of discrete inputs: words, sentences, images, and audio, making semantic similarity and downstream learning feasible (Bengio et al., 2003). Over the past decade, unimodal text embeddings have evolved into multimodal embeddings that jointly encode text and images into a shared space, enabling retrieval, grounding, and cross-modal reasoning (Radford et al., 2021; Li et al., 2023). While this shift has powered significant advances in multimodal understanding, it also raises questions about the stability of these representations under minor, human-preserving edits to the input (Zhou et al., 2024; Liu, 2024).

A key vulnerability is robustness to local lexical variation (Moradi and Samwald, 2021). Substitutions at the character, word, or sentence level can perturb an embedding in ways that mislead a model even when the change preserves meaning for humans. In this work, we focus on word-level synonym substitution: a setting that is intuitively harmless for people yet often misleads models, especially when tokenization, morphology, or pretraining distributions underrepresent particular lexical forms.

Most robustness studies of synonym substitution focus on English (Ribeiro et al., 2020; Morris et al., 2020; Zeng et al., 2021). For Ukrainian, a morphologically rich, lower-resource language, both empirical measures and standardized evaluation tools are scarce, as it ranks in the lower tiers of data availability in CC100 (Conneau et al., 2020) and falls outside the high-resource classes in Joshi et al. (2020). This gap limits ability to diagnose failures and improve multilingual and multimodal systems.

At the same time, many state-of-the-art (SOTA) multimodal encoders that produce embeddings are trained from scratch on web-scale English-centric corpora. Multilingual coverage is frequently bootstrapped via machine translation (dos Santos et al., 2023; Carlsson et al., 2022). This introduces well-known issues: translation artifacts, domain and register mismatches, and lexical biases that distort the target-language signal.

We therefore frame the central problem: there is no reliable measure of synonym-robustness for Ukrainian multimodal embeddings, and the ecosystem lacks the tooling to build one. Concretely, Ukrainian has limited high-quality synonym resources (e.g., WordNet-style inventories with usable synsets), and there are no ready-to-use synonym-substitution methods for robustness testing (candidate generation, form inflection, grammatical filtering).

This paper makes four contributions: (i) We construct a naive dictionary-based synonym substitution method for Ukrainian and quantify how it affects multimodal retrieval performance. (ii) We

*These authors contributed equally to this work

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20458–20468
November 4-9, 2025

develop an LLM-based substitution method that generates context- and grammar-aware Ukrainian synonyms and evaluate its impact. (iii) We introduce a hybrid method that combines dictionary coverage with LLM selection and inflection, yielding higher-quality substitutions and more informative robustness assessments. (iv) We propose and evaluate a fine-tuning approach, using synonym-augmented data, to improve text–to–image retrieval under the Synonym-Substitution Attack (SSA). Finally, we replicate our evaluation and fine-tuning approach on Czech to assess cross-lingual generality and observe consistent gains, indicating transfer beyond a single language.

2 Related Work

Prior work on multilingual and multimodal embeddings follows several distinct strategies that differ in data requirements, computational cost, and language coverage. First, using real paired data, the monolingual CLIP (Radford et al., 2021) pipeline trains a single joint image–text encoder from scratch. It assumes high-quality image–text pairs and significant compute resources; for many languages, including Ukrainian, such resources are limited, making monolingual training impractical.

A second line of work replaces CLIP's original English text encoder with a multilingual language model and trains on multilingual image–text data (Cherti et al., 2023). This approach bootstraps cross-lingual lexical coverage. These models are often "ready-to-use" for many languages out of the box. However, representation quality for a given target language depends on its diversity and quality in the training mixture.

A third approach uses teacher–student transfer to adapt CLIP to new languages (Chen et al., 2023; Carlsson et al., 2022). In such approaches, English CLIP acts as the teacher while a multilingual or target-language encoder learns to replicate its image–text alignment. These methods leverage parallel or translated data and are more compute-efficient than full training, but translation artifacts and lexical biases can distort the embedding space.

Another approach to generate multimodal embeddings is vision–language models (Liu et al., 2024; OpenAI, 2024; Team et al., 2023). However, many of them are closed and do not expose embedding APIs, and they are expensive.

The Ukrainian Visual-Word Sense Disambiguation benchmark (Laba et al., 2024) offers a way to measure multimodal model performance in Ukrainian, showing M-CLIP[1] (Carlsson et al., 2022) as the best baseline. However, OpenCLIP multilingual variants were not evaluated. To fill this gap, we evaluated OpenCLIP models that support Ukrainian on the same benchmark and compared it against M-CLIP. We observed that OpenCLIP model[2] achieved stronger performance metrics on the Ukrainian Visual-WSD benchmark (Table 5), making it the primary model for our study.

Synonym-based adversarial attacks are well-studied in high-resource languages such as English, where various approaches exist to generate semantically and syntactically valid substitutions (Jin et al., 2020; Li et al., 2020). Those methods, however, depend on resources that are either missing or not practically usable in Ukrainian, e.g., WordNet or counter-fitting synonym embeddings (Mrkšić et al., 2016). Even the recently released Wikipedia-mapped Ukrainian WordNet offers only one lemma per synset and thus no usable synonym sets (Romanyshyn et al., 2024). Recently (Mudryi and Ignatenko, 2025) adapt English SSA to Ukrainian via multilingual models, but the approach targets only text and is engineering-heavy. Also, prior work shows that many automatic substitution methods for English still yield grammatically or semantically invalid replacements (Chiang and Lee, 2023). LLM-based synonym substitution (Wang et al., 2024) offers a way to avoid external language resources, but, to the best of our knowledge, it has not been evaluated for low-resource languages.

Given these limitations, instead of directly adapting existing synonym substitution techniques from English, we propose a novel approach leveraging LLMs in combination with an unstructured Ukrainian synonym dictionary to generate contextually and grammatically appropriate synonym replacements for Ukrainian.

3 Methodology

3.1 Problem Definition

We study the robustness of multimodal embeddings to meaning-preserving lexical variation in Ukrainian. Our goal is to quantify and improve the robustness of text–to–image retrieval when a caption is modified by a single-word synonym. We focus on Ukrainian and use English as a high-

[1] XLM-Roberta-Large-Vit-B-16Plus

[2] CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k

resource reference point.

We evaluate text–to–image retrieval only and report HIT@1, HIT@5, and MRR. HIT@k is the proportion of queries whose ground-truth image appears within the top k; Mean Reciprocal Rank (MRR) summarizes overall ranking quality across the full list.

We generate perturbed captions with two SSA methods: (1) Dictionary - synonyms from a Ukrainian list, inflected to match the source word's grammar; (2) LLM (GPT-4o) - contextually generated synonyms, with prompts enforcing meaning preservation and correct inflection.

We conduct all experiments on Multi30K (Elliott et al., 2016), the multilingual extension of Flickr30K (Young et al., 2014) containing 31,014 images, each paired with five crowd-sourced English captions. We use the Ukrainian extension of Multi30K (Saichyshyna et al., 2023), which provides human-translated captions aligned to the same images. We follow the standard train/validation/test splits, and apply synonym substitutions only to test captions.

As shown in Table 1, English outperforms Ukrainian on unperturbed captions, and the gap widens under both substitution methods. With GPT-4o, HIT@1 drops by 11.2% in English vs. 21.2% in Ukrainian ($\approx$ 26% vs. $\approx$ 66% relative). This supports our hypothesis that Ukrainian is more disrupted by synonym substitutions due to richer morphology and lower number of data for pretraining.

SSA method	Lang	HIT@1	HIT@5	MRR
Unperturbed	Ukr	32.1	54.3	42.6
Dictionary	Ukr	7.6	39.3	22.8
GPT-4o	Ukr	10.9	44	26.3
Unperturbed	Eng	41.6	65.7	52.7
Dictionary	Eng	22	52.8	36.2
GPT-4o	Eng	30.4	60.7	44.1

Table 1: OpenCLIP text–to–image retrieval on Multi30K test set for Unperturbed captions and two SSA methods: Dictionary and GPT-4o. Metrics are reported in %.

3.2 Approach

Generating high-quality synonym substitutions for Ukrainian presents challenges due to its rich morphology and inflectional complexity. Ukrainian's rich morphology (number, gender, case) makes meaning-preserving substitutions far harder than in English. Furthermore, the lack of modern SSA frameworks increases the challenge of evaluating model robustness.

To address these issues, we explore three synonym generation approaches. The first method is dictionary-based, using the (Synonimy.info, 2025), which contains 9,200 synonym groups. While this resource offers broad lexical coverage, it has noisy entries (parsing errors, outdated terms, and context-dependent synonyms), and replacements must also match Ukrainian grammar. We partially address this issue with Pymorphy 3 (Halaiko, 2025) to inflect chosen synonyms to the sentence context.

The second method leverages a LLM (GPT-4o (OpenAI, 2024)) to generate context-aware synonyms that are semantically, morphologically and grammatically correct. Our prompt template and few-shot examples illustrating valid, inflectionally correct substitutions are provided in Appendix G. We used GPT-4o because our university provided the necessary budget/credits to support API-based generation at scale.

While LLM-only substitution yields fluent, context-aware replacements, it lacks full coverage of Ukrainian's lexical diversity (rare lemmas, dialectal variants, domain terms). On other hand, dictionary improves coverage, yet entries can be noisy, and in a morphologically rich language like Ukrainian, applying case/number/gender rules often can't be done automatically. Recognizing the strengths and limitations of both methods, we propose a third method, the hybrid approach, in which dictionary candidates are refined by the LLM to select synonyms that best maintain sentence meaning and grammatical integrity.

The hybrid method supplies the LLM with the original sentence and the dictionary-derived candidate list for each detected noun. For every noun, we inject its dictionary synonyms into the prompt and instruct the model to return at most 10 relevant candidates, inflecting each selected synonym to match grammar. The exact hybrid prompt (which reuses the GPT-only template with an added "dictionary candidates" field and the same few-shot examples) is provided in Appendix G.

3.3 Robustness Enhancement

To evaluate whether the issues arise from limited training data, we fine-tune OpenCLIP on the Ukrainian Multi30K training set and explore whether additional in-language data reduce the gap.

Assuming that multilingual models are often trained on machine-translated data and may in-

herit associated biases, we expand the corpus with LLM-generated synonym substitution sentences. For each Multi30K training pair, we create several variants by replacing exactly one caption noun with a context-appropriate, inflected synonym generated by the same LLM method from 3.2; the image remains unchanged. We refer to this extended training setup (Multi30K + synonym-augmented captions) as synonym-augmented fine-tuning. Training continues on the union of the original and augmented pairs using the standard contrastive objective. By adding these meaning-preserving variants, we aim to improve robustness to linguistic shifts in low-resource languages.

We keep most model's layers frozen to retain general multilingual knowledge (Zhai et al., 2022), while selectively unfreezing a subset of text and vision layers for targeted adjustments. The hyperparameters are detailed in Appendix C.

4 Results

SSA Method	HIT@1	HIT@5	MRR
Dictionary	7.6	39.3	22.8
GPT-4o	10.9	44	26.3
Hybrid	16.8	49	31.5
Unperturbed	32.1	54.3	42.6

Table 2: OpenCLIP text–to–image retrieval under three SSA methods, evaluated on the Multi30K Ukrainian test set. Metrics are reported in %.

4.1 SSA results

Table 2 presents the image retrieval performance of OpenCLIP under three different SSA methods, showing substantial variation. A dictionary-based approach yields the lowest scores across all metrics. We attribute this drop in accuracy to the dictionary's limited contextual awareness: while it offers broad lexical coverage, the synonyms it provides often fail to match the morphological and semantic context of each sentence.

In contrast, the GPT-based approach improves performance by generating synonyms that better preserve both meaning and grammatical form, resulting in more relevant synonym substitution.

Finally, the smallest drop in performance comes from the hybrid method. By combining the dictionary's lexical scope with the LLM's context sensitivity, Hybrid yields more precise synonym matches and higher retrieval scores.

Given these results, we asked whether the SSA methods differ in substitution quality. We manually reviewed 500 replacements per method, labeling each as (1) correct (meaning and grammar preserved), (2) correct lemma but grammatically incorrect, or (3) incorrect (alternative meaning). The audit (Table 3) shows that Hybrid replacements are the most valid both grammatically and semantically. During the manual review, we also observed that GPT replacement often generates hyperonyms instead of synonyms or loosely related terms rather than true synonyms, while Hybrid consistently preserves sense and inflection; illustrative examples appear in Appendix F, Table 8.

Following manual analysis, we examined how the model represents synonyms. We visualized attention-based predictions using (Chefer et al., 2021). Consider the caption "Чоловік у синій сорочці лагодить велосипед у жовтій кімнаті" (A man in a blue shirt is fixing a bicycle in a yellow room). With "велосипед" (bicycle), OpenCLIP correctly aligns the token to the corresponding image region; however, when the lower-frequency synonym "ровер" is used, the alignment degrades. The tokenizer splits "ровер" into "ро" and "вер", with the latter receiving minor importance, leading to weaker text–image alignment. A full visualization is provided in Appendix B.

These analyses implicate underrepresented vocabulary; next we test whether fine-tuning, especially with synonym-augmented data, mitigates these errors and improves robustness.

SSA Method	Dict	GPT	Hybrid
Correct	35.2	84.6	**90.4**
Grammar	27.6	2.4	**1.2**
Semantic	37.2	13	**8.4**

Table 3: Manual evaluation of SSA quality across different methods, with results presented as percentages.

4.2 Robustness enhancement results

Fine-tuning on Multi30K demonstrates that even a simple approach can improve text–to–image retrieval under SSA, as presented in Table 4. HIT@1 nearly triples for dictionary substitutions (7.6% → 19.5%), doubles for GPT-4o (10.9% → 23.1%), and rises for hybrid (16.8% → 26.7%), with MRR improving by roughly +10% across all cases. These consistent gains highlight the impact of exposing the model to in-language training data.

Compared to Multi30K-only fine-tuning, synonym-augmented training yields consistent gains: HIT@1 improves slightly for dictionary substitutions (19.5% → 19.7%), more under hybrid (26.7% → 28%) and the most under GPT-4o (23.1% → 25.1%), with MRR increasing by about +0.6 to +1.7%. Across all substitution methods, synonym-augmented fine-tuning consistently delivers the best scores.

Two key insights emerge. First, fine-tuning alone improves text–to–image retrieval under SSA. Second, synonym-augmented fine-tuning generalizes beyond GPT-style substitutions and achieves the best results across all three SSA methods, while maintaining nearly the same performance as Multi30K fine-tuning on the original test set (average metrics drop of -0.19).

Nevertheless, none of the perturbed results approach the strongest unperturbed scores (39.3% HIT@1 and 50.7% MRR from Multi30K fine-tuning), underscoring how disruptive synonym variation remains.

SSA Method	Model	HIT@1	HIT@5	MRR
Unpert.	OpenCLIP	32.1	54.3	42.6
	Multi30K FT	**39.33**	**64.02**	**50.73**
	Synonym FT	39.07	63.76	50.69
Dict	OpenCLIP	7.6	39.3	22.8
	Multi30K FT	19.53	50.95	34.12
	Synonym FT	**19.78**	**51.57**	**34.72**
GPT-4o	OpenCLIP	10.9	44	26.3
	Multi30K FT	23.13	55.64	37.98
	Synonym FT	**25.14**	**56.36**	**39.72**
Hybrid	OpenCLIP	16.8	49	31.5
	Multi30K FT	26.69	58.27	41.19
	Synonym FT	**28.08**	**58.94**	**42.34**

Table 4: Results of OpenCLIP, Multi30K fine-tuned, and synonym-augmented fine-tuned models on the Multi30K Ukrainian test set under unperturbed and three SSA methods. Metrics are reported in %.

4.3 Czech replication

To assess whether our findings are specific to Ukrainian or generalize to other languages with rich morphology, we replicate the study on Czech using the Czech version of Multi30K. Applying the same SSA methods yields the same pattern as in Ukrainian: synonym substitutions degrade text–image retrieval, and the Hybrid method shows the least degradation (Table 7).

The absolute and relative drops under SSA are smaller for the Czech than for the Ukrainian. We hypothesize two reasons: (i) shorter average caption length in the Czech translations (fewer opportunities for substitution-induced mismatch), and (ii) stronger token coverage for Czech in the XLM-RoBERTa vocabulary (Conneau et al., 2020), which reduces harmful subword fragmentation.

We further fine-tune on the Czech portion of Multi30K and observe improvements under all SSA methods, with synonym-augmented fine-tuning achieving the best overall text–to–image retrieval under SSA. As in Ukrainian, synonym-augmented fine-tuning does not degrade performance on the unperturbed captions, indicating gains in robustness without sacrificing original accuracy.

These results provide evidence that our approach is not language-specific: synonym-augmented fine-tuning improves retrieval robustness across multiple Slavic languages. Full Czech results and implementation details appear in Appendix E.

5 Conclusion

This work presents the first systematic evaluation of text–to–image retrieval under SSA for Ukrainian. In the absence of ready-to-use SSA frameworks for this low-resource language, we establish baseline methods using a dictionary-based and an LLM-based (GPT-4o) methods, analyzing their respective strengths and weaknesses. Building on these insights, we propose a novel Hybrid substitution method that better balances lexical coverage with grammatical and semantic correctness.

We further demonstrate that fine-tuning on the Ukrainian Multi30K dataset improves retrieval across all SSA methods. Extending this idea, we introduce synonym-augmented fine-tuning, in which additional training pairs are created by substituting nouns with LLM-generated synonyms. This approach improves text–to–image retrieval under SSA while maintaining performance on the unperturbed data.

Finally, by replicating our experiments on Czech, another morphologically rich Slavic language, we show that both the Hybrid substitution method and synonym-augmented fine-tuning generalize beyond Ukrainian. Together, these results highlight a practical path toward more robust multimodal embeddings in low-resource languages.

Limitations

While our study provides insights into the robustness of multimodal models under synonym substitution in a low-resource language, several limitations remain. First, our synonym substitution strategies rely on predefined lexical resources and large language models, which may introduce biases based on their training data. The dictionary-based method lacks contextual awareness, while the LLM-based approach depends on the quality of its training corpus and may generate synonyms that do not always preserve the intended meaning. Second, while results on two Slavic languages indicate cross-lingual viability, extending the study to non-Slavic morphology (e.g., Turkish, Swahili) and to generation tasks such as VQA remains future work. Third, our experiments keep the original XLM-R tokenizer unchanged; unseen sub-words may therefore still degrade retrieval. Exploring selective vocabulary expansion or adapter-level token embeddings is left for future work.

Ethical Considerations

This study focuses on enhancing the robustness of multimodal retrieval models and does not involve human subjects, personal data, or sensitive information. However, ethical concerns arise regarding potential biases in the synonym substitution methods and the pretraining data of the models used. The dictionary-based approach may contain outdated or culturally specific synonyms that do not reflect contemporary language use. Additionally, large language models, such as GPT-4o, can inherit biases from their training data, which may influence synonym generation and, consequently, model predictions.

We used AI assistants for minor coding assistance, such as debugging and optimizing LaTeX formatting, and for grammar and clarity improvements in writing. However, AI was not used to generate text from scratch, and all research, analysis, and conclusions were solely conducted by the authors.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

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.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 397–406.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mCLIP: Multilingual CLIP via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Cheng-Han Chiang and Hung-yi Lee. 2023. Are synonym substitution attacks really synonym substitution attacks? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1853–1878, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Gabriel Oliveira dos Santos, Diego Alysson Braga Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia Da Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. 2023. CAPIVARA: Cost-efficient approach for improving multilingual CLIP performance on low-resource languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 184–207, Singapore. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

fraug library. 2024. Synonyms dictionaries. Synonym dictionary dataset derived in part from Apache

OpenOffice / Hunspell / OpenOffice thesaurus resources.

Danylo Halaiko. 2025. pymorphy3: Morphological analyzer / inflection engine for Russian and Ukrainian. https://pypi.org/project/pymorphy3/. Version 2.0.4, MIT License.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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.

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. In *International conference on machine learning*, pages 19730–19742. PMLR.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Yizhi Liu. 2024. Comparison of the robustness of multimodal models and unimodal models under text-based adversarial attacks. *Applied and Computational Engineering*, 103:117–122.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

Volodymyr Mudryi and Oleksii Ignatenko. 2025. Precision vs. perturbation: Robustness analysis of synonym attacks in Ukrainian NLP. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 131–146, Vienna, Austria (online). Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o technical report. *OpenAI Technical Report*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. 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.

Synonimy.info. 2025. Словник синонімів української мови (Synonym Dictionary of the Ukrainian Language).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2024. Generating valid and natural adversarial examples with large language models. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1716–1721. IEEE.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78.

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An open-source textual adversarial attack toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371, Online. Association for Computational Linguistics.

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.

Wanqi Zhou, Shuanghao Bai, Qibin Zhao, and Badong Chen. 2024. Revisiting the adversarial robustness of vision language models: a multimodal perspective. *arXiv preprint arXiv:2404.19287*.

Authors' Contributions

Volodymyr Mudryi (VM). Led dictionary-based substitutions (resource cleaning, inflection), generated dictionary candidate lists for the Hybrid approach, ran qualitative error analysis of GPT-only outputs, curated datasets, replicated Czech experiments, formatted tables/figures.

Yurii Laba (YL). Designed LLM-based substitutions (prompts, few-shots), wrote Hybrid prompts and refinement instructions, prepared Multi30K data for fine-tuning, executed both fine-tuning runs and hyperparameter selection, conducted attention-based explainability visualizations, led result aggregation, integrated Ukrainian Visual-WSD baselines, maintained repo/reproducibility.

Both authors jointly defined scope, experimental protocols, and reviewed all results.

A Benchmark Comparison on Ukrainian Visual-WSD

Table 5 reports HIT@1 and MRR on the Ukrainian Visual-WSD benchmark for off-the-shelf multilingual OpenCLIP and M-CLIP checkpoints (no fine-tuning on the benchmark). Because there is a lack of good-quality Ukrainian multimodal benchmarks, we adopt the latest version of Ukrainian Visual-WSD benchmark[3]; its latest release includes 174 unique homonyms. OpenCLIP ViT-H/14 (XLM-R large) attains the best scores, outperforming the strongest M-CLIP variant by ≈ 6 HIT@1 and ≈ 3.9 MRR. These results motivated our choice of the OpenCLIP ViT-H/14 multilingual model in the main experiments.

Model	Variant	HIT@1	MRR
CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	OpenCLIP	**43.82**	**60.6**
XLM-Roberta-Large-Vit-B-16Plus	M-CLIP	37.80	56.7
CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	OpenCLIP	37	54.4
XLM-Roberta-Large-Vit-L-14	M-CLIP	36.48	55.5
LABSE-Vit-L-14	M-CLIP	35.17	53.96
XLM-Roberta-Large-Vit-B-32	M-CLIP	34.65	53.9

Table 5: HIT@1 and MRR metrics comparison of multilingual OpenCLIP versus M-CLIP on the extended Ukrainian Visual-WSD benchmark.

[3] Hugging Face: Ukrainian Visual-WSD

B Attention-based Prediction Visualization

We use the attention-based explainability of Chefer et al. (2021), which propagates relevance through transformer attention to produce cross-modal attribution maps-more architecture-faithful than raw gradients. In Fig. 1, велосипед (bicycle) yields sharp focus on the bike region, while the low-frequency synonym ровер is split into subwords (ро, вер), diffusing token importance and weakening image alignment. The maps expose a concrete failure mode-synonym-induced tokenization shifts degrade grounding-supporting our quantitative drops and motivating synonym-aware augmentation or selective vocabulary expansion.

C Hyperparameters

The OpenCLIP model used in our experiments is based on the ViT-H/14 architecture, which contains approximately 632 million parameters. We conducted both fine-tuning on the Multi30K dataset and fine-tuning with synonym-augmented data on a single NVIDIA RTX 3090 GPU, each taking approximately 10 hours to complete. This setup allowed us to efficiently adapt OpenCLIP for improved robustness while maintaining a manageable computational cost.

Table 6 lists all hyperparameters used in our experiments.

Parameter	Value
`-lr`	1×10^{-6}
`-warmup`	1000
`-batch-size`	64
`-epochs`	20
`-accum-freq`	2
`-wd`	0.01
`-lock-image`	True
`-lock-image-unlocked-groups`	2
`-lock-text`	True
`-lock-text-unlocked-layers`	6
`-pretrained`	`frozen_laion5b_-s13b_b90k`
`-model`	`xlm-roberta-large-ViT-H-14`

Table 6: Hyperparameters used in the experiment.

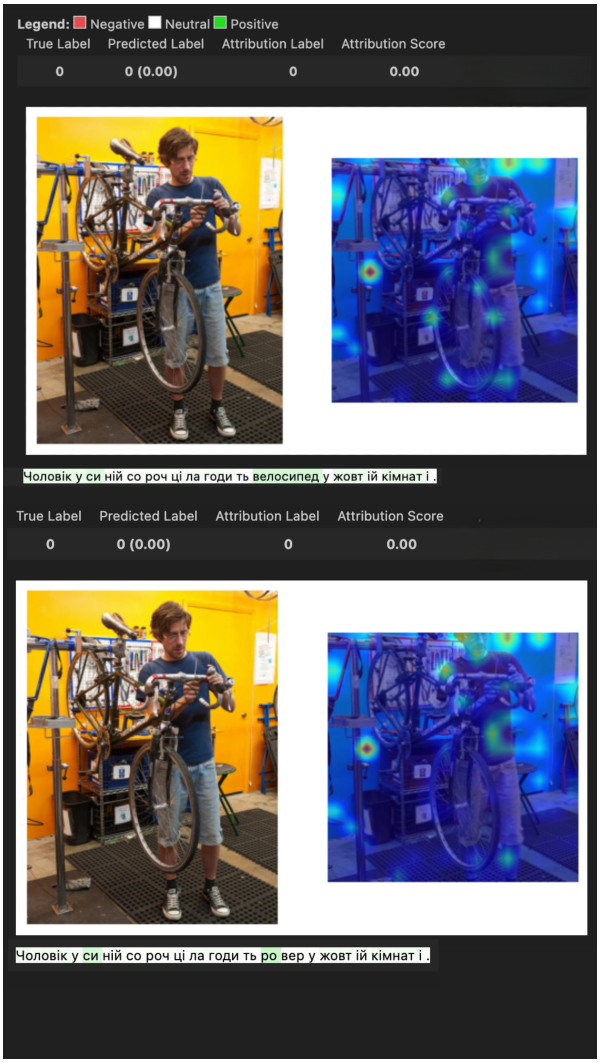

Figure 1: Attribution maps for OpenCLIP image-text alignment. The upper example uses "велосипед" (bicycle), which the model correctly aligns with the image. The lower example replaces it with the synonym "ровер", causing the tokenizer to split it into subwords ("ро" and "вер"), with "вер" receiving no importance, reducing retrieval accuracy.

D Dataset and Model Details

All code, prompts, configs, and checkpoints to reproduce our results are available at github.com/YuriiLaba/UA-B2BE. All the artifacts used in this study are open-source and available for research purposes. The Multi30K dataset (Ukrainian version) is publicly accessible and follows an open-source research license. The OpenCLIP model is released under an Apache 2.0 license, allowing for modification and redistribution within the scope of open research. Any synonym substitution dataset we generate will be made available under a Creative Commons (CC-BY) license to ensure transparency and reproducibility while complying with open-access principles. We fixed the GPT-4o generation parameters to temperature = 0.1 with a controlled random state to ensure reproducibility of synonym outputs.

E Extended Czech Analysis

To replicate our Ukrainian pipeline in Czech, we kept the same data splits, tokenization, image resolution, freezing policy, and evaluation. For the Dictionary setting, we used a following Czech dictionary (fraug library, 2024) and inflected selected lemmas to context with Stanza's morphological analyses (Qi et al., 2020). The GPT-4o prompt was identical to the Ukrainian version but translated to Czech (same constraints and few-shots).

Table 7 reports Czech performance for all SSA methods and fine-tuning variants. Two observations are consistent with the Ukrainian case:

- **Higher baselines.** Czech captions yield stronger results than Ukrainian across most SSA settings. For example, the GPT-4o substitution baseline reaches 15.5 HIT@1 / 48.6 HIT@5, compared to 10.9 / 44.0 for Ukrainian.
- **Largest relative boost for Dictionary.** Synonym-augmented fine-tuning improves HIT@1 from 14.7 $\rightarrow$ 20.1 under Dictionary substitutions (+5.3), a stronger relative gain than in the Hybrid case (21.3 $\rightarrow$ 33.7, +12.4 absolute but smaller relative margin given the higher baseline). This mirrors the Ukrainian trend: fine-tuning closes the robustness gap most when the baseline is weakest.

Overall, the Czech results reinforce our main claim: synonym-based data augmentation consistently improves robustness, with the largest relative benefits under noisy dictionary substitutions. Importantly, synonym-augmented fine-tuning preserves unperturbed accuracy (40.9 HIT@1 vs. 40.7 for plain fine-tuning), showing that robustness gains come without sacrificing clean performance. Together with the Ukrainian study, these findings suggest that synonym-augmented training is a practical strategy for strengthening multimodal encoders across morphologically rich, lower-resource languages.

SSA Method	Model	HIT@1	HIT@5	MRR
Unpert.	OpenCLIP	32.76	54.68	43.19
	Multi30K FT	40.67	64.88	51.98
	Synonym FT	**40.88**	**65.04**	**52.16**
Dict	OpenCLIP	14.73	32.7	23.83
	Multi30K FT	19.07	41.03	29.78
	Synonym FT	**20.07**	**41.83**	**30.61**
GPT-4o	OpenCLIP	15.53	48.63	30.83
	Multi30K FT	29.23	59.1	43.09
	Synonym FT	**29.37**	**59.9**	**43.47**
Hybrid	OpenCLIP	21.33	52.1	35.49
	Multi30K FT	33.27	61.63	46.36
	Synonym FT	**33.67**	**62.27**	**46.71**

Table 7: Results of OpenCLIP, Multi30K fine-tuned, and synonym-augmented fine-tuned models on the Multi30K Czech test set under unperturbed and three synonym-substitution attacks (Dictionary, GPT-4o, Hybrid). Metrics are reported in %.

F Qualitative Error Analysis of GPT-only Substitutions

Frequent failure modes.

- **Low-frequency gap.** GPT struggles to produce dialectal or rare synonyms (e.g. хата 'cottage') that the dictionary or Hybrid methods can surface.
- **Hypernym drift.** It often proposes hypernyms or loosely related words instead of true synonyms, degrading retrieval accuracy (see Table 8).

[4]Brovko, Barbos, and Ryabko are common Ukrainian dog names-metaphoric but still referential to 'dog'.

Word	Src	Synonym (UA/EN)	Count
капелюх (hat)	GPT	шапка (cap)	18
		панама (panama hat)	4
	Hybrid	капелюшок (little hat)	18
		бриль (straw hat)	12
собака (dog)	GPT	друг (friend – hypernym)	1
	Dict	свинюка (swine – insult)	7
	Hybrid	пес (male dog)	7
	Hybrid	псиний (doggish)	6
	Hybrid	цуцик (puppy)	6
	Hybrid	шавка (cur)	4
	Hybrid	бровко (Brovko[4])	3
	Hybrid	барбос (Barbos*)	3
	Hybrid	рябко (Ryabko*)	1
	Hybrid	псина (mongrel)	1

Table 8: Examples showing GPT's hypernym drift vs. Hybrid's synonym quality (500-sample audit).

G Synonym-Substitution Prompt (SSA)

System message

```
You are an assistant receiving a sentence in English.

Your task:
  1. Identify ALL nouns.
  2. Generate up to five relevant synonyms for each noun, if that many exist.
  3. For each noun and each of its synonyms:
     - Replace **only** this noun (in all its repetitions and forms) in the original sentence with the corresponding synonym, preserving the meaning.
     - If a correct grammatical case is needed, adjust the synonym accordingly.
  4. In the returned JSON, display each noun and its synonym in the **nominative case**; in the "new_sentence" field, show the sentence with the noun replaced.
  5. Return the result in JSON format, WITHOUT any additional text, numbering, or lists.

The JSON format should be as follows:

    "results": [
    "noun": "example of a noun in the nominative case",
    "synonym": "corresponding synonym in the nominative case",
    "new_sentence": "sentence with the noun replaced"
    ,
    ...
    ]

Do not add any text outside this JSON, and do not use '1.', '2.', '-', or similar markers.
Return only valid JSON that can be parsed without errors.
```

User message

```
Here is an example:

Example sentence: "The student was browsing a textbook in the library."
Example output:

    "results": [
    "noun": "student",
    "synonym": "pupil",
    "new_sentence": "The pupil was browsing a textbook in the library."
    ,
    "noun": "textbook",
    "synonym": "manual",
    "new_sentence": "The student was browsing a manual in the library."
    ,
    "noun": "library",
    "synonym": "reading room",
    "new_sentence": "The student was browsing a textbook in the reading room."

    ]

Now perform the same task for this sentence:
"INPUT_SENTENCE"

Remember: no additional text or numbering. Only valid JSON with the 'results' key.
```