UkrMeme: Multimodal Explainable Persuasion Detection in Ukrainian Memes

Nazar Andrushko[1], Vera Schmitt[2,3,4], Veronika Solopova[2,3]

[1]Ukrainian Catholic University, Lviv, Ukraine
[2]Johannes Gutenberg University Mainz, Mainz, Germany
[3]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[4]Centre for European Research in Trusted AI (CERTAIN)

nazar.checkemail@ucu.edu.ua vera.schmitt@uni-mainz.de veronika.solopova@uni-mainz.de

Abstract

Memes play a central role in modern online communication, especially in contexts of political conflict where visual and textual cues combine to convey persuasive messages. We study persuasive technique detection in Ukrainian memes using vision–language models (VLMs). We introduce UkrMeme, a dataset of 200 organically produced Ukrainian wartime memes annotated for persuasion techniques, and PTM-UA, a Ukrainian translation of a subset of the Propaganda 950 benchmark that preserves original visual content while translating meme text. We evaluate VLMs in a zero-shot prompting setup across image, text, and multimodal inputs using hierarchical F1 metrics. Finally, we compare Layer-wise Relevance Propagation (LRP) and model-generated token attributions and analyze the semantic categories of evidence emphasized by each explanation method.

1 Introduction

Memes have become one of the central communicative form in modern information ecosystem, combining visual and textual signals into compact multimodal artifacts. In contexts of political conflict and war, memes often serve as weapon of soft propaganda, especially easy to plant for foreign audiences, but it can also be instrument of collective sense-making, identity construction, and psychological coping (Liagusha and Iarovyi, 2025; of Excellence, 2015; Mejova et al., 2025). Since the beginning of the full-scale invasion of Ukraine in 2022, Ukrainian online spaces have witnessed an unprecedented proliferation of war-related memes. Scholars and journalists have noted that Ukrainian meme culture functions as a form of grassroots resilience, black humor, and communal morale-building rather than merely as an instrument of coordinated information warfare (Reuters Institute for the Study of Journalism, 2023; Bilaniuk, 2023; Bondarenko, 2025; Bouchard and Antsybor, 2025).

While detection of persuasion techniques in multimodal content has recently received increased attention in Natural Language Processing (NLP), most existing benchmarks are English-centric and focus on identifying persuasive techniques in political messaging (Dimitrov et al., 2021). Cross-lingual and culturally grounded analyses remain limited, especially for low-resource languages and contexts of active conflict. Moreover, the majority of prior work emphasizes model performance, with comparatively less attention to faithfulness and explainability of model decisions.

In this work, we investigate how large language models (LLMs) and vision-language models (VLMs) identify such techniques in memes in Ukrainian. Importantly, while we use persuasion labels from propaganda studies, this study does not assume that Ukrainian memes constitute state-coordinated propaganda, but a form of grassroots resilience. Hence, we treat persuasion techniques as an analytical lens for examining how persuasive rhetoric and emotional framing emerge organically in grassroots digital creativity during wartime. Our focus is on understanding both model behavior and the rhetorical structure of Ukrainian meme discourse. We construct two complementary datasets: (i) we translate into Ukrainian a subset of a widely used multimodal persuasive meme benchmark (Dimitrov et al., 2021), preserving images while translating meme text into Ukrainian. This controlled setup allows us to isolate language effects while holding visual content constant; (ii) we compile and annotate a novel dataset of 200 organically produced Ukrainian wartime memes, curated from public online sources. Each meme is independently annotated by two native Ukrainian speakers using a multi-label taxonomy, followed by adjudication to ensure consistency. Figure 1 shows a representative example from the dataset.

We evaluate multimodal VLM models (Qwen3-VL-4B-Instruct (Qwen Team, 2025) and 8B ver-

Proceedings of the Fifth Workshop on NLP for Positive Impact (NLP4PI), pages 530–543
October 28, 2026 ©2026 Association for Computational Linguistics

Figure 1: Example from the UKRMEME dataset. Gold labels: *Smears*, *Loaded Language*. en: "A monument to "Vladimir Lenin founding Ukraine" bronze, granite, Alzheimer's"

sion) and GPT-4.1-mini and Gemma 3 4B in a zero-shot prompting setup on both datasets. Beyond standard micro- and macro-averaged F1 scores, we conduct a faithfulness-oriented explainability analysis. Specifically, we compare token-level Layer-wise Relevance Propagation (LRP) (Bach et al., 2015a; Montavon et al., 2019) saliency with model based token attributions. We quantify overlap between these signals and analyze divergences across datasets and modalities. Finally, we offer a pilot empirical perspective on how grassroots wartime meme culture in Ukraine employs persuasive techniques as part of civic resilience.

Both datasets and code is available in GitHub[1].

2 Related Work

Automatic detection of persuasion techniques has been widely studied in textual political discourse. Early work focused on news articles and political speeches, introducing fine-grained taxonomies of persuasive techniques and corresponding supervised benchmarks (Da San Martino et al., 2019a; Dmitriev et al., 2019). The labels and annotation quality for this task has been disputed as subjective and usually low-annotator agreement is expected (Sahitaj et al., 2025). With the rise of multimodal communication, attention has shifted toward propaganda in internet memes, where meaning emerges from the interaction between text and image (Dimitrov et al., 2021; Kmainasi et al., 2025; Singh et al., 2025). These works demonstrate much lower performance than purely textual propaganda detection. At the same time, multimodal models were shown to outperform text-only systems on meme classification tasks, particularly when OCR-extracted text is incorporated into the visual pipeline. Notably, multimodal task inherited the labels from early propaganda works (Da San Martino et al., 2019b), with small image and cross-modality based label additions. However, existing benchmarks remain predominantly English-centric, while other languages and cross-lingual transfer in memes remain underexplored, especially in culturally specific contexts such as wartime meme production.

To our knowledge, no prior work systematically evaluates persuasion technique detection in Ukrainian memes under a prompt-only setting or compares text-only and vision-language models using faithfulness-oriented explainability metrics.

3 Methods

3.1 Data

We utilize two distinct datasets for our experiments: a translated subset of the Propaganda 950 dataset from Dimitrov et al. (2021), PTM-UA (Propaganda Techniques in Meme-Ukrainian) and a newly collected dataset of Ukrainian memes, UKRMEME.

3.1.1 Translated Propaganda 950 (PTM-UA)

To obtain a Ukrainian-language evaluation set, we translate the test split of the Propaganda 950 dataset (Dimitrov et al., 2021) using a semi-automated pipeline. The process consists of four stages: (1) text detection using EasyOCR[2], (2) removal of detected text via LaMa inpainting (Suvorov et al., 2022)[3], (3) translation of the extracted English text into Ukrainian with GPT-5 (OpenAI, 2025), prompted to preserve rhetorical intent while producing concise, colloquial output, and (4) rendering the translated text back onto the image using automatic font sizing, text wrapping, and adaptive color selection for readability.

After the automated pipeline, we manually review all 200 translated test samples. We discard samples where the text placement was unsatisfactory (e.g., text overflowing bounding boxes or overlapping with important visual elements), where the inpainting process removed crucial visual information, or where the translation was inaccurate. The majority of rejected samples suffered from poor text placement, with a smaller number lost to inpainting artifacts and very few due to translation errors. After filtering, we retain **158 samples** for evaluation. The label distribution for the translated

[1] https://github.com/eingrid/persuasion-techniques-ua-memes

[2] https://github.com/JaidedAI/EasyOCR

[3] https://github.com/enesmsahin/simple-lama-inpainting

Figure 2: Example from PTM-UA: original English meme (left) and its Ukrainian translation (right). Text is inpainted and re-rendered while preserving the visual content. Gold labels: *Appeal to Authority*, *Smears*.

dataset is shown in Figure 4. Note that although the original Propaganda 950 dataset does not include an explicit No Propaganda label, a small number of test samples had none of the 22 propaganda techniques identified; we treat these as No Propaganda.

An example of the translation pipeline output is shown in Figure 2.

3.1.2 UKRMEME

To complement the translated data with natively Ukrainian content, we collect 200 memes from open Ukrainian-language sources, including Telegram channels and the SUCHO Meme Wall[4]. The images were collected in early 2026 and represent a variety of persuasive and non-persuasive content circulating in Ukrainian online spaces.

Two trained native-speaking annotators independently label each image using the same 22 technique labels as in the Propaganda 950 dataset (Dimitrov et al., 2021), plus a No Propaganda label[5] for images that do not employ any persuasion technique. Disagreements are resolved by a designated curator. The label distribution (Figure 3) shows a substantially larger share of No Propaganda samples and a different technique ranking compared to PTM-UA (Figure 4), reflecting the distinct nature of Ukrainian wartime meme content.

We report inter-annotator agreement following the multi-label methodology of Marchal et al. (2022). Table 1 presents set-level measures; Table 2 reports per-class Cohen's κ for classes with at least 5 occurrences from either annotator. The MASI-weighted (Passonneau, 2006) Krippendorff's α is **0.406** and the mean per-class κ is **0.404**, indicating moderate agreement. Agreement is highest for *No Propaganda* ($\kappa = 0.776$); among techniques, *Smears* ($\kappa = 0.511$) and *Appeal to Fear/Prejudice* ($\kappa = 0.506$) show the strongest agreement. Low κ for classes such as *Loaded Language* ($\kappa = 0.149$) reflects very low prevalence rather than poor annotation quality. Prior work attributes low agreement in this task to the inherent subjectivity of persuasion technique annotation and treats IAA as a signal for annotator management rather than a strict quality threshold (Sahitaj et al., 2025).

Measure	Observed	Expected	Adjusted
Boot-match	0.764	0.303	0.661
Boot-precision	0.628	0.184	0.544
Boot-recall	0.698	0.227	0.610
Boot-F1	0.643	0.180	0.565
Augmented-κ	0.535	0.238	0.390
MASI Krippendorff's α: **0.406**			

Table 1: Set-level inter-annotator agreement for UKRMEME.

Label	κ
No Propaganda	0.776
Smears	0.511
Appeal to Fear/Prejudice	0.506
Exaggeration/Minimisation	0.415
Causal Oversimplification	0.409
Glittering Generalities (Virtue)	0.394
Flag-waving	0.358
Name Calling/Labeling	0.338
Appeal to (Strong) Emotions	0.305
Transfer	0.279
Loaded Language	0.149
Mean	**0.404**

Table 2: Per-class Cohen's κ for UKRMEME labels with at least 5 occurrences from either annotator.

3.2 Experimental set-up

In this section, we describe our approach to persuasion detection and the explainability analysis we conduct using two different methods: Layer-wise Relevance Propagation (LRP) (Bach et al., 2015b) and VLM-based token attribution.

3.2.1 Motivation

LRP and VLM-based token attribution represent fundamentally different explainability approaches. LRP propagates relevance scores backward through the network, assigning importance based on each token's mathematical contribution to the output, a model-internal, gradient-based explanation. However, LRP explanations depend strongly on the

[4] https://memes.sucho.org/

[5] While we use "No Propaganda" label inheriting from previous datasets in the field, we underline that we do not assume that the rest of the dataset constitutes propaganda, but instead samples of grassroots resilience memes.

Figure 3: Label distribution for the UKRMEME dataset (200 samples).

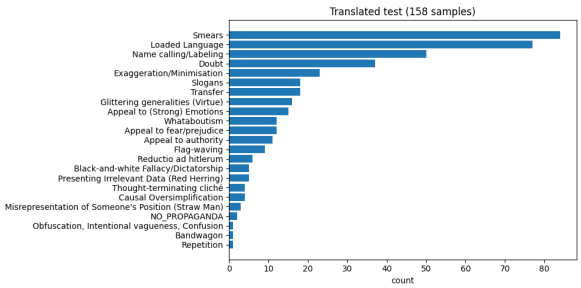


Figure 4: Label distribution for the translated PTM-UA (158 samples).

chosen propagation rules and model architecture, meaning that analysis may reflect methodological choices rather than the model's true decision process. Token attribution for Language models relies on the model's own reasoning to identify important tokens as part of its generation, which is a linguistic, model-external explanation. We hypothesize that comparing these complementary explanation methods can reveal systematic differences in the evidence emphasized by each approach.

3.2.2 Models

We evaluate four models across two tiers. **Qwen3-VL-4B-Instruct** and **Qwen3-VL-8B-Instruct** (Qwen Team, 2025) are used for both persuasion detection and explainability analysis: as open-weight models, they provide full access to internal activations, enabling LRP alongside VLM-based token attribution on the same model. Comparing the two sizes additionally reveals scale effects on detection and explanation quality. We also report detection-only results for two additional models: **GPT-4.1-mini** and **Gemma 3 4B**. These are excluded from the explainability analysis for different reasons detailed in Appendix A.

3.2.3 Detection Metrics

We report hierarchical F1 following (Kiritchenko et al., 2006; Dimitrov et al., 2024), which awards partial credit when predicted and gold labels share ancestors in the taxonomy. The taxonomy organizes 22 techniques under categories such as Ethos, Pathos, Logos, and Ad Hominem (Dimitrov et al., 2024). Each label set is extended to include all ancestor nodes except the root (Persuasion), and precision, recall, and F1 are computed over these extended sets. We report three variants: micro-averaged (the official SemEval metric), macro-averaged over taxonomy nodes, and mean per-sample F1. Full definitions are provided in Appendix D.

3.2.4 Baselines and translation control

We evaluate a weighted-random baseline to contextualize model performance. For each persuasion label, a binary prediction is sampled independently according to its empirical positive frequency in the dataset, and results are averaged over 30 runs. We report the same hierarchical micro- and macro-F1 metrics as for the evaluated models.

Additionally, to assess whether performance on PTM-UA is affected by translation into Ukrainian, we evaluate all four models on the original English versions of the same 158 Propaganda 950 samples retained in PTM-UA after filtering. We use identical input modalities and prompting settings, enabling a paired comparison in which meme identity and visual content are held constant while the textual condition differs between the original English and translated Ukrainian versions.

3.2.5 Explanation Setup

Since our datasets consist of visual content (images containing text), we evaluate across three input modes: (i) *image*, where only the image is provided and the model must read text from it; (ii) *image+text*, where both the image and its extracted text are provided in the prompt; and (iii) *text*, where only the extracted text is provided without the image. This allows us to assess how input modality affects both detection performance and the resulting explanations.

Prompting. We use two zero-shot prompts: one for the image mode and another shared by the image+text and text modes. Both prompts provide the full set of persuasion technique classes and instruct the model to return a JSON object with three fields: `labels` (predicted persuasion techniques), `referred_words` (individual words from the meme text most relevant to the classification), and `referred_image_concepts` (up to five visual

elements supporting the classification, ranked by importance). Full prompts are provided in Appendix B. While the prompt also elicits image-level concept descriptions, we restrict our explainability analysis to textual token attributions due to the difficulty of systematically comparing free-text concept outputs with LRP spatial heatmaps.

LRP extraction. We use `lxt.efficient` with `zennit` to monkey-patch the linear, gated-MLP, RMSNorm, LayerNorm, and attention layers of the Qwen3-VL text decoder with AttnLRP propagation rules (Achtibat et al., 2024) (library defaults, unmodified), replacing plain gradients in the backward pass. For each sample, we re-run a forward pass from the cached generated token sequence, with gradients enabled on the input embeddings, and set the propagation target as the sum of logits at the predicted label token(s). Relevance is the embedding times its gradient, summed over the embedding dimension, then normalized per sample by its max absolute value (non-finite values clipped to zero). For *image+text*, this backward pass uses text tokens only; visual features are not re-injected. To obtain word-level scores, we align the OCR text with the decoded prompt via tokenizer offsets, assign each overlapping subword token to the word containing its span midpoint, and average absolute relevance per word. Words are filtered (stopwords, punctuation, length 1), and the number of LRP words kept per sample is capped to match the count of words in the model's own verbalized explanation for that sample (max $k = 5$), for comparison.

Word-level comparison. We compare VLM-attributed and LRP-selected words using Jaccard similarity to assess the agreement between the two explanation methods.

Category analysis. To characterize the types of explanations each method produces, we categorize tokens into five categories: ENTITY (named people, places, organizations, brands), EMOTIONAL (emotionally loaded or fear/hate/outrage-inducing language), RHETORIC (calls to action, imperatives, slogans, persuasive framing), CONTENT (neutral factual nouns, events, dates, statistics), and OTHER (numbers, punctuation artifacts, unclear items) using GPT-4o (OpenAI, 2024). We then compare the category distributions across methods. The categorization prompt is provided in Appendix C, and its accuracy was validated on a manually annotated sample.

4 Results

We report hierarchical F1 scores across two datasets (PTM-UA and UkrMeme), three input modalities (text, image, image+text) and four models. We also include a weighted-random baseline as a lower bound. For each label, this baseline draws a binary prediction independently with probability equal to that label's positive frequency in the training data, averaged over 30 runs. On PTM-UA it reaches hierarchical micro F1 of 0.376 ± 0.024 and macro F1 of 0.149 ± 0.018, with micro precision 0.503 ± 0.030 and recall 0.300 ± 0.021. On UkrMeme it reaches micro F1 of 0.256 ± 0.025 and macro F1 of 0.107 ± 0.016, with micro precision 0.257 ± 0.024 and recall 0.256 ± 0.026. Both are well below every model result in Table 3, showing that model predictions carry real signal beyond label-frequency priors. The gap between the two baselines is itself telling: PTM-UA's baseline is precision-heavy while UkrMeme's is roughly balanced, the same precision-recall asymmetry we observe for the models themselves below, and a sign of UkrMeme's flatter, No-Propaganda-heavy label distribution compared to PTM-UA. Qwen3-VL-4B and Qwen3-VL-8B are evaluated on both detection and explainability. GPT-4.1-mini and Gemma 3 4B serve as detection-only baselines (see §3.2.2). Table 3 compares all four models on hF1 only, and Table 6 in Appendix B details other metrics for the Qwen models.

4.1 Performance and Cross-Dataset Gap

On PTM-UA, the best micro hF1 is 0.537 (Qwen3-VL-4B, image+text) with per-sample F1 of 0.489. On UkrMeme, performance drops substantially: micro F1 peaks at 0.342 (Qwen3-VL-8B, image+text) and per-sample F1 at 0.388. Macro F1 remains below 0.30 on PTM-UA and below 0.20 on UkrMeme, reflecting poor coverage of infrequent techniques. This persistent gap stems from three factors: (i) UkrMeme's large share of *No Propaganda* samples, which the model frequently mislabels; (ii) the translated memes originate from English-language discourse well represented in pretraining, whereas UkrMeme requires Ukrainian wartime and cultural knowledge; and (iii) precision-recall profiles are inverted, PTM-UA is precision-oriented (P = 0.569 vs. R = 0.508 for 4B image+text), while UkrMeme is recall-oriented (P = 0.272 vs. R = 0.396), indicating systematic over-prediction on culturally unfamiliar content.

Dataset	Averaging	Mode	hF1			
			Qwen3-VL-4B	Qwen3-VL-8B	GPT-4.1-mini†	Gemma 3 4B†
PTM-UA	Micro	text	0.440	0.416	0.556	0.498
		image	0.483	**0.524**	0.592	0.549
		image+text	**0.537**	0.523	**0.613**	**0.569**
	Macro	text	0.195	0.198	0.255	0.255
		image	0.219	0.257	0.350	0.297
		image+text	**0.245**	**0.291**	**0.356**	**0.309**
	Per-sample	text	0.367	0.351	0.517	0.449
		image	0.437	**0.480**	0.549	0.531
		image+text	**0.489**	0.475	**0.577**	**0.541**
UkrMeme	Micro	text	0.305	0.326	0.343	0.288
		image	**0.323**	0.337	0.379	0.324
		image+text	0.323	**0.342**	**0.384**	**0.334**
	Macro	text	0.125	0.156	0.160	0.131
		image	0.142	0.183	0.192	0.166
		image+text	**0.143**	**0.193**	**0.198**	**0.168**
	Per-sample	text	**0.277**	**0.388**	0.235	0.191
		image	0.256	0.274	0.258	0.246
		image+text	0.258	0.276	**0.270**	**0.251**

Table 3: hF1 comparison across all four models on PTM-UA (n=158) and UkrMeme (n=200). Best hF1 per averaging method is **bolded** within each dataset and averaging type. †GPT-4.1-mini and Gemma 3 4B are detection-only baselines; explainability analysis is not conducted for these models (see §3.2.2).

The two additional baselines reinforce these trends. GPT-4.1-mini achieves the highest micro hF1 across both datasets: 0.613 on PTM-UA (image+text) and 0.384 on UkrMeme (image+text), outperforming both Qwen models on detection. Gemma 3 4B lags behind GPT-4.1-mini but exceeds the Qwen models on PTM-UA recall-heavy metrics (micro hF1 0.569 image+text), driven by aggressive over-prediction (high recall, lower precision). On UkrMeme, Gemma 3 4B's performance is similar to Qwen3-VL-4B (micro hF1 0.334 vs. 0.323). The cross-dataset gap likely reflects several interacting factors, including UkrMeme's larger share of No Propaganda samples, differences in label distribution, and the greater cultural and contextual specificity of native wartime memes. As examined in §4.2, translation into Ukrainian additionally affects the Qwen models, although this effect is almost negligible in GPT-4.1-mini and Gemma 3 4B.

4.2 Language Effect: Original vs. Translated English

To isolate the effect of Ukrainian translation from dataset difficulty, we additionally run all four models on the *original English* versions of the same 158 PTM-UA memes retained after filtering (§3.1.1), using identical prompts and modalities, and compare against their translated-set hF1 from Table 3.

Model	Mode	Orig. EN Mi / Ma hF1	Translated Mi / Ma hF1
GPT-4.1-mini	text	0.548 / 0.270	0.556 / 0.255
	image	0.615 / 0.362	0.592 / 0.350
	image+text	0.603 / 0.347	0.612 / 0.356
Gemma 3 4B	text	0.495 / 0.257	0.498 / 0.255
	image	0.542 / 0.309	0.549 / 0.297
	image+text	0.566 / 0.318	0.569 / 0.314
Qwen3-VL-4B	text	0.482 / 0.214	0.440 / 0.195
	image	0.578 / 0.280	0.483 / 0.219
	image+text	0.577 / 0.308	0.537 / 0.245
Qwen3-VL-8B	text	0.479 / 0.238	0.416 / 0.198
	image	0.553 / 0.326	0.524 / 0.257
	image+text	0.548 / 0.338	0.523 / 0.291

Table 4: Original English vs. Ukrainian-translated PTM-UA, hierarchical micro/macro F1, on the 158 retained samples. Translated-condition values for Qwen3-VL-4B/8B are taken from Table 3; Mi = micro hF1, Ma = macro hF1.

Table 4 reports micro and macro hF1 for both conditions.

For GPT-4.1-mini and Gemma 3 4B, performance is broadly similar between original and translated conditions (differences within ± 0.03 hF1 across all modes), indicating translation itself introduces little additional degradation beyond the source dataset's inherent difficulty. The Qwen models show a more consistent gap in the opposite direction: both 4B and 8B score higher on original

English than on translated Ukrainian across every mode (e.g. 4B image+text: 0.577 vs. 0.537; 8B image+text: 0.548 vs. 0.523), suggesting that the Qwen models are comparatively more sensitive to the translated Ukrainian condition.

4.3 Effect of Input Modality

On PTM-UA, image+text yields the best F1 for the 4B model across all averaging methods, confirming that supplying extracted text reduces OCR errors. The 8B model's image-only performance is competitive with image+text (micro 0.524 vs. 0.523), suggesting its stronger OCR partially closes the gap. Text-only consistently lags behind, indicating visual context carries non-trivial signal.

On UkrMeme, modality effects diverge by metric. Image and image+text modes achieve higher micro F1 through improved recall, but text-only yields the best per-sample F1 for both 4B (0.277) and 8B (0.388). The image modes over-predict labels, with higher recall coming at the cost of substantially lower precision, penalizing individual samples. The model interprets visual cues in wartime memes as additional evidence for persuasion techniques, leading to systematic over-labeling.

4.4 Effect of Model Scale

Scaling from 4B to 8B yields uneven gains. On PTM-UA, the 8B model improves image-only micro F1 (0.524 vs. 0.483) and achieves the best macro F1 (0.291), indicating better rare-technique coverage. On UkrMeme, the largest gain appears in text-only per-sample F1 (0.388 vs. 0.277, a 40% relative improvement). In multimodal modes, gains are modest on both datasets.

4.5 No Propaganda Label Ablation

The original Propaganda 950 taxonomy does not include a *No Propaganda* label; we add it to the prompt as an explicit option. In PTM-UA only 2 samples genuinely lack annotations, yet both models predict *No Propaganda* for 31 samples in text-only mode. Removing this option (Table 7 in Appendix B) yields the largest gains in text-only mode; image and image+text improvements are marginal. The 4B model is more responsive: of the 31 text-only *No Propaganda* predictions, only 8 become empty when the label is removed, with the rest replaced by technique labels. The 8B model is more conservative (22 of 31 become empty rather than committing to a label).

4.6 Per-Class Confusion Patterns

Figures 5 and 6 show confusion matrices for the 4B model across modalities. Cell (i, j) is the raw count of samples where gold label i and predicted label j both occur: letting $G \in \{0,1\}^{N\times K}$ and $P \in \{0,1\}^{N\times K}$ be the binary gold and predicted label indicator matrices over N samples and K labels, the matrix is $G^\top P$; no row- or column-wise normalization is applied.[6] Figure 9 in Appendix shows PTM-UA without the *No Propaganda* label.

Across both datasets, *Loaded Language* is the dominant attractor class: on PTM-UA it accumulates 283 column pairs against 214 row pairs, and on UkrMeme 179 column pairs against only 12 row pairs. The model defaults to this semantically broad label when uncertain. *Smears* and *Name Calling/Labeling* show strong mutual confusion on both datasets, consistent with their semantic proximity. Several mid-frequency classes (*Black-and-white Fallacy*, *Causal Oversimplification*, *Glittering Generalities*, *Transfer*) never appear on the diagonal in either dataset.

The key cross-dataset divergence concerns *No Propaganda*: on PTM-UA this row contains only 4 label pairs, whereas on UkrMeme it contains 170 and drives the majority of false positives, 36 pairing with predicted *Loaded Language*, 31 with *Slogans*, and 29 with *Exaggeration/Minimisation*. The model lacks a reliable mechanism for abstaining on emotionally charged but non-propagandistic wartime content.

4.7 LRP and token attribution

Dataset related differences. Across the four Qwen settings, the token-level explanations show a clear split between the translated and native Ukrainian meme conditions. In the translated dataset, both VLM-based token attribution and LRP are dominated by tokens tied to the original English political discourse (e.g. "trump", "democrats"). This pattern is stable for both text+image and text-only settings, suggesting that translation preserves the main lexical persuasion cues but also keeps the dataset anchored in the source-domain. In contrast, the native Ukrainian dataset is much more centered on the local wartime and sociopolitical context, with frequent tokens such as "ukrainians", "missiles", "drones", "Russia", "Putin", "alert", and "Tomahawk".

[6] Since the task is multi-label, each cell counts gold-prediction label pairs; a sample with multiple gold and predicted labels contributes to several cells.

Figure 5: Confusion matrix for UkrMeme (Qwen3-VL-4B, image+text). The model massively over-predicts *Loaded Language* and *Slogans*, particularly for gold *No Propaganda* samples.

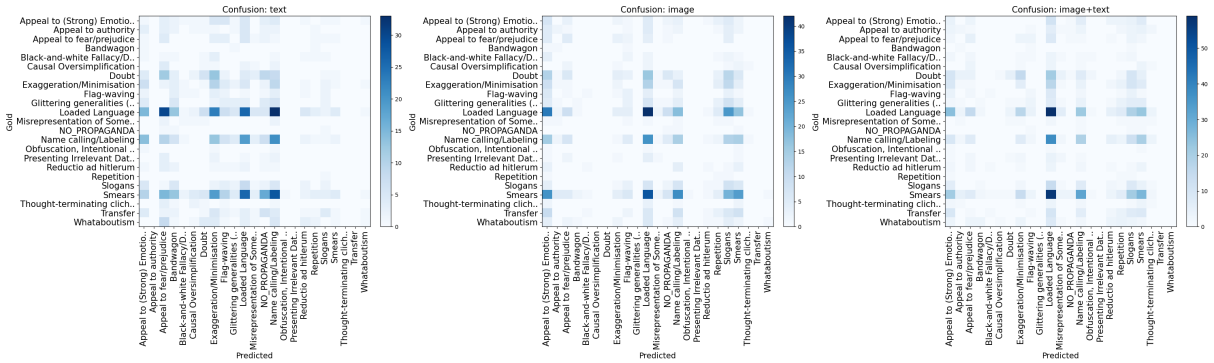
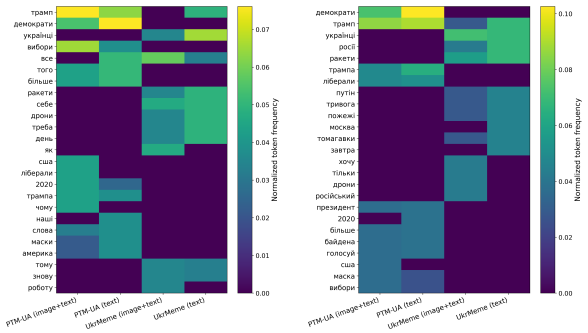


Figure 6: Confusion matrix for PTM-UA (Qwen3-VL-4B, image+text). *Loaded Language* acts as the dominant attractor class.



(a) LRP attributions (b) VLM attributions

Figure 7: Token attribution distributions across dataset settings. Columns correspond to PTM-UA (translated memes) and UkrMeme (native Ukrainian memes), with both multimodal (image+text) and text-only variants.

Explanation agreement varies across dataset settings. Native multimodal memes exhibit the highest median overlap between VLM explanations and LRP attributions, suggesting more stable evidence selection. In contrast, translated memes show lower and more variable agreement, indicating that translation artifacts and multimodal context may destabilize token-level attribution alignment.

Modality differences. The comparison between text+image and text-only also reveals an important modality effect. In the translated condition, the two modes are fairly similar, which suggests that much of the signal is recoverable from OCR text alone. In the native Ukrainian condition, however, the text+image model produces a broader and less purely lexical set of salient tokens, including more context-dependent tokens like "watch", "people", "works", "self", and "day", whereas the text-only model concentrates more on explicit textual triggers such as "missiles", "Trump", "Russia", and "Putin." This indicates that for native memes, the visual channel contributes additional contextual evidence that is not reducible to extracted text.

Low accuracy settings. The low-hF1 (< 0.2) examples are especially informative. A recurring pattern is that VLM-based token attribution and LRP often overlap strongly even when the prediction is wrong, with several cases showing very high Jaccard agreement, even up to 1.0. That means the model is often consistently focusing on the same cues across explanation methods, but still mapping them to the wrong persuasion label. In other words,

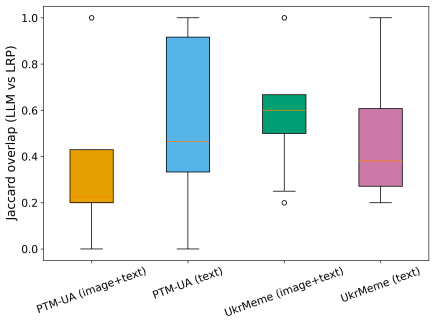


Figure 8: Agreement between VLM-based and LRP token attributions explanations for low-hF1 samples across dataset settings. The distribution shows moderate overlap across conditions, with higher median agreement for the UkrMeme multimodal setting.

	PTM-UA			UkrMeme		
	Sh.	**VLM**	**LRP**	**Sh.**	**VLM**	**LRP**
Entity	22.9	15.8	19.4	20.5	13.0	16.9
Emotional	33.1	33.7	12.7	25.9	32.4	12.7
Rhetoric	17.1	22.7	19.4	15.5	17.6	15.9
Content	23.4	26.4	34.3	33.2	30.9	33.1
Other	3.6	1.1	14.2	4.9	6.2	21.3
n	363	273	268	425	324	314

Table 5: Category distribution (%) of attributed tokens by source. Total word counts shown as n.

many failures are not due to random attention, but due to semantic confusion between nearby categories such as Smears, Loaded Language, Name-calling, Slogans, or Appeal to emotions. This is important because it suggests that the main weakness is not cue detection itself, but label discrimination.

At the same time, there are also low-hF1 cases with low overlap between VLM and LRP tokens, especially in some translated text-only examples. Those cases likely reflect OCR noise, translation artifacts, or unstable lexical grounding, where the model's stated rationale and its internal relevance distribution diverge. This seems particularly visible in examples involving odd lexical fragments, borrowed English tokens, or noisy meme text.

4.8 Category Analysis of Attributed tokens

We categorize all attributed tokens into five semantic categories (ENTITY, EMOTIONAL, RHETORIC, CONTENT, OTHER; see Section 3) using GPT-4o (OpenAI, 2024) with a structured categorization prompt (Appendix C), validated at 80.3% accuracy on a manually annotated sample of 61 tokens. We partition tokens into *shared* (selected by both methods), *VLM-only*, and *LRP-only* groups (Table 5).

The two methods select qualitatively different evidence. VLM-attributed tokens are disproportionately EMOTIONAL (33.7% on PTM-UA, 32.4% on UkrMeme) and RHETORIC (22.7%, 17.6%), suggesting that the model's verbalized reasoning gravitates toward affectively loaded and persuasive-sounding language when justifying its prediction.

LRP-only tokens show a different profile: CONTENT dominates ($\sim$34% on both datasets), while EMOTIONAL drops to 12.7%. The OTHER category is also elevated (14.2% on PTM-UA, 21.3% on UkrMeme), reflecting LRP's tendency to assign relevance to function words and punctuation that carry mathematical but not semantic importance.

Shared tokens have the highest ENTITY proportion ($\sim$21-23%) and lowest OTHER ($\sim$4-5%), indicating that when both methods converge, they agree on named entities, specific people, organizations, and places, as anchors for classification. These patterns are consistent across datasets despite strong differences in content and language, with the main divergence being a higher OTHER ratio in UkrMeme's LRP-only group, likely reflecting noisier OCR output in native Ukrainian memes.

5 Discussion and Conclusion

Our results highlight both the potential and limitations of current vision–language models for analyzing multimodal political discourse in low-resource and culturally specific contexts. While models achieve moderate performance on translated memes derived from English benchmarks, performance drops substantially on native Ukrainian memes, suggesting that cultural knowledge and contextual grounding remain key bottlenecks. Input modality also plays a complex role: multimodal inputs improve recall but often lead to systematic over-prediction of persuasion techniques, particularly in emotionally charged wartime content. Increasing model scale yields uneven gains, with improvements primarily in text-only settings rather than multimodal reasoning. Finally, our explainability analysis reveals a consistent divergence between model-based token attribution reasoning and LRP-based attributions: verbal explanations emphasize emotional and rhetorical cues, while internal relevance signals focus more on content words. Together, these findings underscore the challenges of reliable persuasion detection and interpretation in real-world multimodal settings.

Limitations

Our study has several limitations. First, both datasets are relatively small: PTM-UA contains 158 translated memes and UkrMeme 200 native Ukrainian memes. While sufficient for exploratory evaluation, this scale limits statistical power and will not capture the full diversity of Ukrainian meme culture. Second, persuasion technique annotation is inherently subjective, and although inter-annotator agreement is moderate, disagreement reflects genuine ambiguity in rhetorical interpretation rather than purely annotation error, which is, however, conforming with reported agreement problem in prior work.

Third, the translated PTM-UA dataset preserves visual content from English memes while replacing text with Ukrainian translations. While this design isolates language effects, it may introduce translation artifacts and does not fully reproduce authentic Ukrainian meme discourse. Although originally motivated as a possible route toward data augmentation, our results suggest that simple translation-based augmentation is unlikely to provide substantial benefits. Moreover, the labels, coming from propaganda research, may not capture all of the discourse methods genuine civic society creation contains. Future work should analyse the memes in partnership with social scientists and linguists to identify genre-specific phenomena.

Fourth, our experiments rely on zero-shot prompting and therefore do not evaluate supervised or fine-tuned systems. Fifth, explainability analysis is restricted to the two Qwen3-VL models: GPT-4.1-mini is excluded because LRP requires access to model-internal activations unavailable through a closed API, and Gemma 3 4B was excluded because running LRP over a 4B-parameter model for the full dataset was not computationally feasible within the scope of this study. Finally, our explainability analysis focuses on textual token attribution because visual concept explanations are difficult to compare systematically across methods. Future work should extend faithfulness analysis to spatial visual explanations, larger datasets, and open-weight models of comparable size to the closed-API baselines.

Ethical Considerations

This work studies persuasion techniques in memes created during an ongoing armed conflict. Importantly, we do not assume that Ukrainian memes constitute coordinated propaganda; rather, we treat persuasion techniques as an analytical framework for examining rhetorical strategies in grassroots digital communication. Ukrainian wartime memes often serve functions of humor, resilience, and collective coping rather than strategic information warfare. While empirically our annotation procedure showed that persuasion techniques as a set of labels provided a sufficiently expressive annotation framework encoded in UkrMeme dataset, with consideration of the sensitivity of the current war context, we recognize that some readers can find utilization of propaganda and persuasion techniques insulting.

All memes in the UkrMeme dataset were collected from publicly accessible online sources. The dataset is used exclusively for research purposes and does not include private or personally identifiable information. Nevertheless, memes may contain politically sensitive or emotionally charged content. We therefore frame our analysis as descriptive rather than normative and caution against using automated persuasion detection systems as tools for censorship or political moderation. We acknowledge the broader risk that such research can be misused to suppress legitimate political expression. Our goal is to improve transparency and understanding of the phenomena, not to enable automated policing of political speech.

6 Acknowledgment

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101285890. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015a. On pixel-wise explanations

for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015b. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140.

Laada Bilaniuk. 2023. Memes as antibodies: Humor and resilience in ukraine. In *Language in the Time of War*. Routledge.

Ievgeniia Bondarenko. 2025. Black humour in ukrainian war memes: cognitive issues. *The European Journal of Humour Research*, 13(2):41–62.

Michel Bouchard and Daria Antsybor. 2025. Memeing war: the use of humor for hope, resistance, and forging the nation. *National Identities*, 0(0):1–34.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Dimitar Petrov, and Preslav Nakov. 2019a. Fine-grained analysis of propaganda in news articles. In *Proceedings of EMNLP*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Denis Dmitriev and 1 others. 2019. A survey on propaganda detection. In *Proceedings of RANLP*.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A. Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Advances in Artificial Intelligence*, volume 4013 of *Lecture Notes in Computer Science*, pages 395–406. Springer.

Mohamed Bayan Kmainasi, Abul Hasnat, Md Arid Hasan, Ali Ezzat Shahroor, and Firoj Alam. 2025. MemeIntel: Explainable detection of propagandistic and hateful memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30263–30279, Suzhou, China. Association for Computational Linguistics.

Anton Liagusha and Dmytro Iarovyi. 2025. Memes, freedom, and resilience to information disorders: Information warfare between democracies and autocracies. *Social Sciences Humanities Open*, 11:101247.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yelena Mejova, Arthur Capozzi, Corrado Monti, and Gianmarco De Francisci Morales. 2025. Narratives of war: Ukrainian memetic warfare on twitter. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: An overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 193–209.

NATO Strategic Communications Centre of Excellence. 2015. It's time to embrace memetic warfare. Defining memetic warfare as narrative and idea competition akin to propaganda.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2025. GPT-5 Large Language Model. Model version/snapshot from 07 Aug 2025.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*.

Qwen Team. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Reuters Institute for the Study of Journalism. 2023. Memes and morale: decoding ukraine's comedy arsenal against disinformation.

Ariana Sahitaj, Premtim Sahitaj, Veronika Solopova, Jiaao Li, Sebastian Möller, and Vera Schmitt. 2025. Hybrid annotation for propaganda detection: Integrating LLM pre-annotations with human intelligence. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 215–228, Vienna, Austria. Association for Computational Linguistics.

Yashashvi Singh, Adwait Kesharwani, Animesh Chaturvedi, and Rajesh Sharma. 2025. Memeperil: a diverse multi-modal harmful memes dataset for the evolution analytics and llm prompts. *Discover Data*, 3.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with Fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

A Model exclusion from Explainability Study

GPT-4.1-mini is a closed API model, and its internal activations are not accessible, making LRP technically infeasible. Gemma 3 4B is open-weight, but running LRP over a 4B-parameter model with the full dataset was not computationally feasible within the scope of this study. Both models are therefore included as detection baselines only.

B Prompt Templates

We use two prompt templates depending on input modality. Both prompts receive the full list of 22 persuasion technique labels plus `NO_PROPAGANDA` (when included). For the ablation without *No Propaganda* (Section 4.5), we remove the `NO_PROPAGANDA` label and the clause *"or 'NO_PROPAGANDA' if the content doesn't contain persuasion techniques."* All outputs are requested in Ukrainian.

B.1 Image+Text and Text-Only Prompt

```
You are an expert in analyzing communication
strategies in visual media.
Classify the provided content into one or
multiple of the following persuasion techniques:
{unique_labels}
As output, provide JSON with three fields:
"labels", "referred_words", and
"referred_image_concepts".
- "labels": list of persuasion techniques used,
  or "NO_PROPAGANDA" if the content doesn't
  contain persuasion techniques.
- "referred_words": list of the most important
  INDIVIDUAL words from the content text
  supporting your classification. Only single
  words, not phrases. Only key words most
  relevant to the classification.
- "referred_image_concepts": up to 5 visual
  elements most strongly supporting your
  classification, ranked by importance. No
  duplicates. Empty list if no image. The list
  could be empty if the image doesn't provide
  any clues.
Visual elements rules:
- Be concrete: name actual people, objects,
  symbols. Do NOT write abstract scene
  descriptions like "people in a stressful
  situation".
- For visual elements: use the shortest
  identifying phrase -- "Obama"/"Обама",
  "flag"/"прапор". Do not use words on the
  image themselves.
Language note: All values in referred_words and
referred_image_concepts must be in Ukrainian.
JSON format output example:
{"labels": ["label1", "label2"],
 "referred_words": ["word1", "word2"],
 "referred_image_concepts": ["concept1",
 "concept2"]}
Text from the content: {text}
Output:
```

B.2 Image-Only Prompt

```
You are an expert in analyzing communication
strategies in visual media.
Classify the provided content IMAGE into one or
multiple of the following persuasion techniques:
{unique_labels}
As output, provide JSON with two fields:
"labels" and "referred_image_concepts".
- "labels": list of persuasion techniques used,
  or "NO_PROPAGANDA" if the content doesn't
  contain persuasion techniques.
- "referred_image_concepts": up to 5 visual
  elements most strongly supporting your
  classification, ranked by importance. No
  duplicates. The list could be empty if the
  image doesn't provide any clues.
Visual elements rules:
- Be concrete: name actual people, objects,
  symbols. Do NOT write abstract scene
  descriptions like "people in a stressful
  situation".
- For words/text visible in the image: do not
  include ANY of them to
  referred_image_concepts, as they belong to
  the text modality. Only visual elements
  should be included.
- For visual elements: use the shortest
  identifying phrase -- "Obama"/"Обама",
  "flag"/"прапор". Do not use words on the
  image themselves.
Language note: All values in
referred_image_concepts must be in Ukrainian.
JSON format output example:
{"labels": ["label1", "label2"],
 "referred_image_concepts": ["concept1",
 "concept2"]}
Output:
```

B.3 Propaganda Technique Labels

The following 23 labels (including `NO_PROPAGANDA`) are provided to the model as `{unique_labels}`: Appeal to (Strong) Emotions, Appeal to Authority, Appeal to Fear/Prejudice, Bandwagon, Black-and-white Fallacy/Dictatorship, Causal Oversimplification,

Dataset	Averaging	Mode	Qwen3-VL-4B			Qwen3-VL-8B		
			hP	hR	hF1	hP	hR	hF1
PTM-UA	Micro	text	0.538	0.372	0.440	0.492	0.360	0.416
		image	0.570	0.419	0.483	0.536	0.512	**0.524**
		image+text	0.569	0.508	**0.537**	0.519	0.526	0.523
	Macro	text	0.308	0.175	0.195	0.328	0.210	0.198
		image	0.302	0.249	0.219	0.334	0.295	0.257
		image+text	0.280	0.277	**0.245**	0.354	0.320	**0.291**
	Per-sample	text	0.436	0.339	0.367	0.408	0.332	0.351
		image	0.512	0.411	0.437	0.514	0.479	**0.480**
		image+text	0.538	0.480	**0.489**	0.492	0.492	0.475
UkrMeme	Micro	text	0.272	0.346	0.305	0.343	0.310	0.326
		image	0.271	0.399	**0.323**	0.284	0.413	0.337
		image+text	0.272	0.396	0.323	0.279	0.442	**0.342**
	Macro	text	0.123	0.193	0.125	0.187	0.190	0.156
		image	0.150	0.209	0.142	0.195	0.289	0.183
		image+text	0.183	0.196	**0.143**	0.204	0.276	**0.193**
	Per-sample	text	0.294	0.277	**0.277**	0.408	0.385	**0.388**
		image	0.264	0.266	0.256	0.276	0.296	0.274
		image+text	0.267	0.267	0.258	0.269	0.306	0.276

Table 6: Hierarchical F1 results on PTM-UA (n=158) and UkrMeme (n=200) for Qwen3-VL models. Best hF1 per averaging method is **bolded** within each model and dataset.

Averaging	Mode	Qwen3-VL-4B			Qwen3-VL-8B		
		w/ NP	w/o NP	Δ	w/ NP	w/o NP	Δ
Micro	text	0.440	**0.475**	+.035	0.416	**0.432**	+.016
	image	0.483	**0.527**	+.044	**0.524**	0.515	−.009
	image+text	0.537	**0.550**	+.013	0.523	**0.529**	+.006
Macro	text	0.195	**0.214**	+.019	0.198	**0.211**	+.013
	image	0.219	**0.229**	+.010	0.257	**0.276**	+.019
	image+text	0.245	**0.249**	+.004	0.291	**0.301**	+.010
Per-sample	text	0.367	**0.427**	+.060	0.351	**0.368**	+.017
	image	0.437	**0.482**	+.045	**0.480**	0.470	−.010
	image+text	0.489	**0.511**	+.022	0.475	**0.485**	+.010

Table 7: PTM-UA results with and without the *No Propaganda* label (n=158). The largest improvements appear in text-only mode, where *No Propaganda* is most frequently predicted.

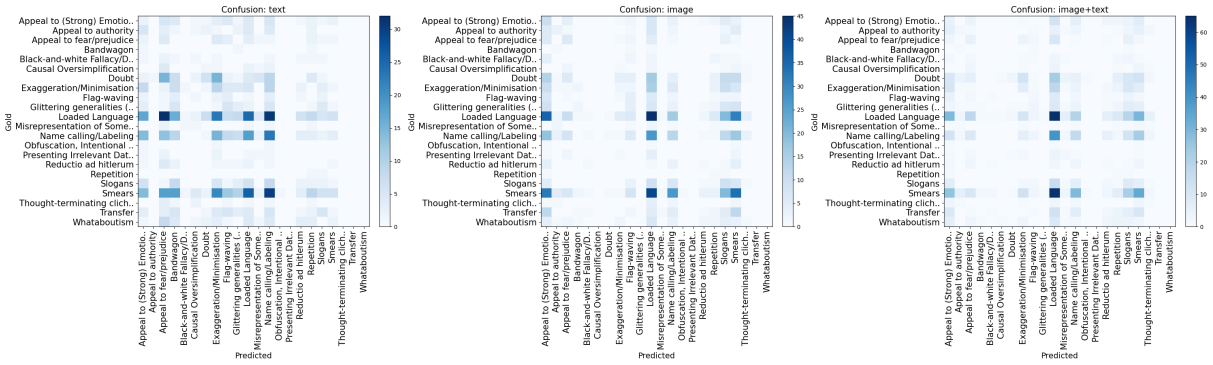


Figure 9: Confusion matrix for PTM-UA without *No Propaganda* (Qwen3-VL-4B, image+text). Former abstentions are redistributed into partially correct technique labels.

Doubt, Exaggeration/Minimisation, Flag-waving, Glittering Generalities (Virtue), Loaded Language, Misrepresentation of Someone's Position (Straw Man), Name Calling/Labeling, Obfuscation/Intentional Vagueness/Confusion, Presenting Irrelevant Data (Red Herring), Reductio ad

Hitlerum, Repetition, Slogans, Smears, Thought-terminating Cliché, Transfer, Whataboutism, and NO_PROPAGANDA.

C Category Annotation Prompt

The following prompt is used with GPT-4o to categorize attributed words. The placeholders `{meme_text}`, `{pred_labels}`, and `{word_list}` are filled per sample.

You are analyzing memes. The meme text is in Ukrainian. Categorize words in context.
Meme text: "{meme_text}"
Predicted persuasion techniques: {pred_labels}
The following words were flagged as important in this meme by automated methods:
{word_list}
For each word, assign ONE category based on its role in THIS specific meme:
- ENTITY: named people, places, organizations, brands
- EMOTIONAL: emotionally loaded, fear/hate/outrage-inducing language
- RHETORIC: calls to action, imperatives, slogans, persuasive framing
- CONTENT: neutral factual nouns, events, dates, statistics
- OTHER: numbers alone, punctuation artifacts, unclear

Return ONLY a JSON object mapping each word to its category.
No markdown, no explanation.
Example: {"word1": "ENTITY", "word2": "EMOTIONAL", "word3": "CONTENT"}

D Hierarchical F1 Definitions

Given a taxonomy tree T, let $\hat{Y}_i$ and Y_i denote the predicted and gold label sets for sample i, and let $\uparrow Y$ denote the extension of Y to include all ancestor nodes in T (excluding the root).

Per-sample hierarchical precision and recall:

$$hP_i = \frac{|\uparrow\hat{Y}_i \cap \uparrow Y_i|}{|\uparrow\hat{Y}_i|}, \quad hR_i = \frac{|\uparrow\hat{Y}_i \cap \uparrow Y_i|}{|\uparrow Y_i|} \tag{1}$$

Per-sample hierarchical F1:

$$hF1_i = \frac{2 \cdot hP_i \cdot hR_i}{hP_i + hR_i} \tag{2}$$

Mean per-sample hierarchical F1:

$$\overline{hF1} = \frac{1}{N}\sum_{i=1}^{N} hF1_i \tag{3}$$

Micro-averaged hierarchical F1 (official SemEval metric):

$$\begin{aligned} hP_\mu &= \frac{\sum_{i=1}^{N} |\uparrow\hat{Y}_i \cap \uparrow Y_i|}{\sum_{i=1}^{N} |\uparrow\hat{Y}_i|}, \\ hR_\mu &= \frac{\sum_{i=1}^{N} |\uparrow\hat{Y}_i \cap \uparrow Y_i|}{\sum_{i=1}^{N} |\uparrow Y_i|}. \end{aligned} \tag{4}$$

$$hF1_\mu = \frac{2 \cdot hP_\mu \cdot hR_\mu}{hP_\mu + hR_\mu} \tag{5}$$

Macro per-label hierarchical F1. For each node n in the hierarchy $\mathcal{N}$ (all leaves and internal nodes, excluding root), define binary indicators:

$$y_{i,n} = 1[n \in \uparrow Y_i], \quad \hat{y}_{i,n} = 1[n \in \uparrow\hat{Y}_i] \tag{6}$$

Per-node precision, recall, and F1:

$$\begin{aligned} P_n &= \frac{\sum_{i=1}^{N} y_{i,n}\hat{y}_{i,n}}{\sum_{i=1}^{N} \hat{y}_{i,n}}, \\ R_n &= \frac{\sum_{i=1}^{N} y_{i,n}\hat{y}_{i,n}}{\sum_{i=1}^{N} y_{i,n}}, \\ F1_n &= \frac{2P_nR_n}{P_n + R_n}. \end{aligned} \tag{7}$$

Macro average over active nodes $\mathcal{N}^* = \{n \in \mathcal{N} : \sum_i y_{i,n} > 0 \vee \sum_i \hat{y}_{i,n} > 0\}$:

$$hF1_M = \frac{1}{|\mathcal{N}^*|}\sum_{n \in \mathcal{N}^*} F1_n \tag{8}$$