MONETA: Multimodal Industry Classification through Geographic Information with Multi Agent Systems

Anonymous ACL submission

Abstract

Industry classification schemes are integral parts of public and corporate databases as they classify businesses based on economic activity. Due to the size of the company registers, manual annotation is costly, and fine-tuning models with every update in industry classification schemes requires significant data collection. We replicate the manual expert verification by using existing or easily retrievable multimodal resources for industry classification. We present MONETA, the first multimodal industry classification benchmark that contains text sources (Website, Wikipedia, Wikidata) and geospatial information paralleling company addresses (OpenStreetMap and satellite imagery). Our dataset enlists 1,000 businesses in Europe with 20 economic activity labels according to EU guidelines (NACE). Our training-free and resource-agnostic Multimodal Large Language Model (MLLM) baseline reaches 62.10% accuracy. Our results indicate an increase of up to 22.80% with the combination of multi-turn design, context enrichment, and classification explanations. We will release our dataset and the enhanced guidelines.

1 Introduction

Geospatial finance (Gopal and Pitts, 2024) is an emergent and complex field that links financial and economic attributes to environmental and spatial resources. Due to the inclusion of both text and image resources, geospatial finance tasks are multimodal as part of AI research. One of the important milestones in multimodal research, Multimodal Large Language Models (MLLM) such as Llava (Liu et al., 2023b,a, 2024), InternVL (Chen et al., 2024; Wang et al., 2024b), and QwenVL (Bai et al., 2023; Wang et al., 2024a; Bai et al., 2025), can utilize geospatial content in addition to other sources to analyze and validate economic attributes. Traditionally, machine learning research developed unimodal automatic industry classification, a research sub area that can benefit from geospatial sources, (Werb et al., 2024). We propose, MONETA, a multimodal industry classification benchmark. On this task, Figure 1, we link text and geospatial resources to the economic activities using Zero-Shot and Multi-Turn pipelines.

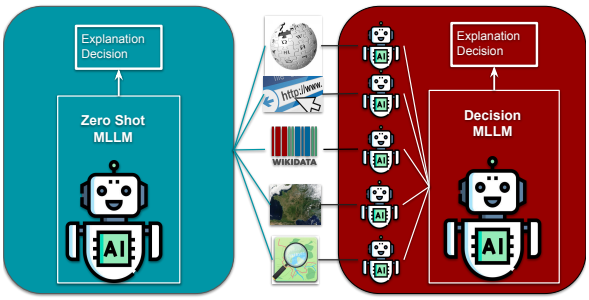

Figure 1: Zero-Shot vs Multi-Turn comparison for MONETA. (1) Zero-shot pipeline (left): Available resources are forwarded into the industry classifier MLLM together. Explanations and classifications are obtained. (2) Multi-Turn pipeline (right): Each resource is processed by separate specialized agents. Intermediate *clues* from these agents are processed by the decision-making agent, returning explanation and classification.

Industry classification schemes date back to the 1930s and aim to classify entities based on economic activity. Due to the ever-changing nature of businesses with time and location, several systems globally (ISIC (UN, 2008)) and region-specific (NAICS (Ambler and Kristoff, 1998), NACE (European Commission, 2008)) have been proposed. Existing research on automatic industry classification from company recordings, financial reports, and websites (Kühnemann et al., 2020; Béchara et al., 2022; Rizinski et al., 2023; Faria and Seimandi, 2023; Vamvourellis et al., 2024; Malashin et al., 2024; Guo et al., 2025; Dzuyo et al., 2025) has two main drawbacks. First, these approaches use a single modality (text). Geospatial information can substitute in the absence of texts. Second, they fine-tune models that require large datasets and limit them to a single classification scheme.

We connect economic activities to spatial extent to answer following research questions:

- RQ-1: Can MLLMs use geospatial information as well as text for industry classification?
- RQ-2: Which configuration (classification explanations, context enrichment, multi-agent) is more helpful?
- RQ-3: How can we quantify intermediate agent performance with respect to the final prediction and the ground truth labels?

To analyze MLLM performance, we propose a novel task: **Multimodal Industry Classification with Geospatial Information**. In this study, we introduce:

- **MONETA: Multimodal Industry Classification Benchmark** for Europe following NACE (European Commission, 2008) scheme. Our dataset contains 1,000 businesses for 20 sections with OpenStreetMap (OSM) and Satellite images for visual inputs. For each company, we also include at least one external resource of Wikidata, Wikipedia, and website.
- **Multimodal Industry Classification**: An expert domain task for multimodal AI rooted in geospatial finance. We propose Zero Shot and Multi-Turn (Multi-Agent) approaches to utilize various multimodal resources. Our pipeline supports various output configurations and prompting strategies.
- **Novel Intermediate Agent Evaluation:** We provide quantitative measures for final inference certainty. Also, we propose a novel keyword-based strategy to analyze intermediate agent performance with respect to ground truth and decision-making agent prediction.

2 Related Work

2.1 Industry Classification

Industry classification has existed since the 1930s with the Standard Industry Classification (SIC). It provides insight into market and sustainability analysis (Ambrois et al., 2023; Croce et al., 2024). For organizations, classifying businesses accordingly and automatically is an important research domain, (Kühnemann et al., 2020; Faria and Seimandi, 2023). To adapt to the emergence of new fields, various classification schemas have been proposed over time. MSCI and Standard & Poor's (S&P) developed the Global Industry Classification Standard (GICS[1]) while the United Nations Statistics Division released the International Standard Industrial Classification (ISIC) (UN, 2008). ISIC has region-specific variants: North American Industry Classification System (NAICS) (Ambler and Kristoff, 1998) and Statistical Classification of Economic Activities in the European Community (NACE) (European Commission, 2008).

European Union uses NACE, a hierarchical classification scheme built upon ISIC. In NACE, sections, corresponding to the major economic activity group, are denoted with the 21 letters from A to U (Example, *C: Manufacturing*). Sections are followed by 88 divisions, 272 groups, and 514 classes.

For automatic industry classification, Kühnemann et al. (2020) used websites for NACE classes. Béchara et al. (2022) annotated an extract of the World Bank collection for Ghana and South Africa and fine-tuned transformers for ISIC scheme. Rizinski et al. (2023), Faria and Seimandi (2023) and Vamvourellis et al. (2024) used company descriptions to classify industries. Jagrič and Herman (2024) fine-tuned transformers on website data for a custom classification scheme. Guo et al. (2025) fine-tuned separate sentence transformers for various hierarchical levels of NAICS. Dzuyo et al. (2025) used adapters to train LLMs over financial statements for SIC. Malashin et al. (2024) employed a genetic algorithm approach for hyperparameter tuning for NACE's divisions.

Many of the studies above, except Rizinski et al. (2023), relied on fine-tuning which requires extensive data collection and annotation. As fine-tuned models are classification scheme specific, models and datasets have to be updated after every revision. Furthermore, all the studies incorporated unimodal text sources, such as financial statements, which may not be available or accessible for newly-founded companies. Werb et al. (2024) argue that economic activity analysis can be enriched with other sources such as satellite imagery, which is the research gap this study covers.

2.2 Geospatial Understanding

Geospatial AI (GeoAI) methodologies cover a variety of tasks such as Geolocation, (Song et al., 2025; Mendes et al., 2024), Geocoding (Nakatani et al., 2025), Remote Sensing (Tao et al., 2025), Ques-

[1] https://www.msci.com/indexes/index-resources/gics

Dataset	Classes	Samples	Text Source	Image Source	Industry Scheme
Remote Sensing Classification Benchmarks					
UC Merced (Yang and Newsam, 2010)	21	2,100	✗	Satellite	✗
AID (Xia et al., 2017)	30	10,000	✗	Satellite	✗
AID++ (Jin et al., 2018)	46	400,000+	✗	Satellite	✗
CLRS (Li et al., 2020)	25	15,000+	✗	Satellite	✗
Industry Classification Benchmarks					
Dutch Businesses (Kühnemann et al., 2020)	111	40,796	Websites	✗	NACE
GHAZAF (Béchara et al., 2022)	56	∼6,500	Survey text	✗	ISIC
SIRENE (Faria and Seimandi, 2023)	732	∼10 Million	Company Descriptions	✗	NACE
WRDS (Rizinski et al., 2023)	11	34,338	Company Descriptions	✗	GICS
SEC 10K (Vamvourellis et al., 2024)	11 (66)	2,590	Company Descriptions	✗	GICS
Industry Websites (Jagrič and Herman, 2024)	13	66,886	Website	✗	Custom
Economic Activity Records (Malashin et al., 2024)	20 (88)	∼20 Million	Company Descriptions	✗	NACE
SEC EDGAR (Dzuyo et al., 2025)	8	9,582	Financial reports	✗	SIC
ExioNAICS (Guo et al., 2025)	20 (1,114)	20,850	Descriptions + emissions	✗	NAICS
Our Dataset					
MONETA (OURS)	**20**	**1,000**	**Website + Wikipedia + Wikidata**	**Satellite + OSM**	**NACE**

Table 1: Comparison of datasets across remote sensing and industry classification tasks. ✗ indicates the absence of that modality or label scheme.

tion Answering and Fact Verification (Norouzi and Hitzler, 2025; Anderson et al., 2025; Khan et al., 2025), Geospatial Foundation Model and Agents (Mansourian and Oucheikh, 2024; Xu et al., 2024).

Geospatial subtasks, such as remote sensing, aim to extract location and geospatial features from the available resources, for instance, satellite imagery, street views, and OpenStreetMap (OSM[2]). OSM can point to other sources such as Wikipedia, Wikidata, and Websites, from which Norouzi and Hitzler (2025) used Wikidata for geolocation.

Remote sensing datasets are abundant in the AI community, one of which is UC-MED (Yang and Newsam, 2010) with 2,100 images for 21 land use classes. A comprehensive study of annotated remote sensing images, AID (Xia et al., 2017), tested 10,000 images for 30 aerial scene classes with object detectors. Its extension, AID++ (Jin et al., 2018), listed more than 400,000 images and semi-automatically annotated 46 classes. Another important benchmark focusing on continual learning is CLRS (Li et al., 2020). It has 15,000+ images and 25 classes obtained from the category network.

Tao et al. (2025) used these geospatial features in Graph-Based Multimodal Data Fusion to identify Urban Functional Zones (UFZ). Another fusion approach (Chen et al., 2025) applied contrastive learning to understand latent urban features. A historical analysis of poverty mappings in Mongolia (Yang et al., 2024), demonstrated the connection between economic activities and geospatial resources.

Although GeoAI has uncovered spatial economic attributes, none of the existing works have targeted entity-level industry classification. Their lack of industry classification adoption is clearly visible in Table 1. Therefore, our work is the first to investigate the validity of geospatial resources for the industry classification task for businesses.

3 MONETA

We introduce MONETA, a novel multimodal benchmark for industry classification based on NACE. In this section, we explain our mapping and dataset. In the mapping stage, we uncover our novel mapping of OSM tags and NACE sections. We provide an overview of our data sources and distribution.

Mapping: We aim to select OSM entries in Europe for NACE scheme. As there is no direct mapping connecting these two domains, we propose a **novel NACE to OSM** mapping shown in Figure 2.

We first instructed Gemini to generate OSM tags for NACE sections using official guidelines in RDF/XML format. However, due to the task complexity, initial mapping was prone to errors. We integrated a human-in-the-loop setup. We annotated and reconstructed the mapping with GPT and Gemini. In the end, we obtained a list of OSM tags per NACE sections. We added additional filters for data quality: name, address, and external resources (Website, Wikipedia and Wikidata). After querying OSM and grouping the entries by NACE sections, we obtained our gold dataset. We provided more details and examples in Appendix A.1.

Gold Dataset: We sampled 50 entries for each NACE section. Using bounding boxes, we calculated dynamic zoom ratios. To retrieve satellite image, we used the ESRI REST API[3]. ESRI and

[2]https://www.openstreetmap.org/

[3]https://developers.arcgis.com/rest/static-basemap-tiles/

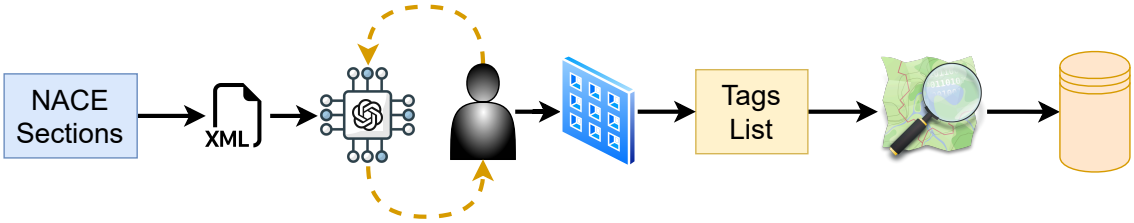

Figure 2: Overview of the dataset preparation process. (1) NACE section XMLs are initially converted to the OSM tags and manually checked by the authors. (2) We added custom filters for data quality and queried Europe OSM data with tag list. (3) Samples are grouped by NACE codes to form the gold dataset.

OSM map service returns static tiles given coordinates and zoom ratio. By concatenating the tiles, we have two images (OSM and satellite) of the same region. We obtained external sources and formed the first multimodal industry classification benchmark, as shown in Table 1.

Our dataset contains 1,000 entries with OSM Tags, two image sources, and at least one external source from Wikidata, Wikipedia, or Website. The data format, distribution and categories are discussed in Appendix, Table 5, and Figures 6 and 7. Entities have NACE sections (A-U) as class labels, except T, which corresponds to households.

4 Experiments

Multimodal industry classification task aims to test MLLM using various resources to predict economic activities. We developed two pipelines for this task that will utilize `MONETA`'s resources: Zero-Shot and Multi-Turn. Zero-Shot pipeline uses the multimodal inputs to select the NACE section in a single inference. Multi-turn pipeline has clue extracting agents for each input type, and a decision-making agent processes these clues.

We then define several experiment dimensions to test with model selection, prompting strategies, input configurations, and output structures. To process intermediate agent responses quantitatively, we define a clue analysis stage incorporating frequency vectors for NACE sections. In our experiments, we propose new metrics to analyze model uncertainty and intermediate agent responses based on frequency vectors.

4.1 Pipeline

In this study, we tested two adaptable and training-free pipelines to accommodate future changes in classification schemes: Zero-Shot and Multi-Turn.

Zero-Shot: We provide inputs with various configurations and instruct MLLM to utilize them to classify entities based on NACE sections.

Multi-Turn: In multi-turn, we have two stages: Clue Extraction and Decision Making. The clue extraction uses agents designed to process input types such as OSM images and website texts. For each entity, we have clue extractor agents up to the number of available input types. Decision-making agent uses intermediate agent responses, *clues*, and the entity name to choose NACE sections.

4.2 Experiment Dimensions

Models: We selected: InternVL 2.5 (1B, 4B, 38B) and InternVL 3 (8B, 14B, 38B, 78B) (Chen et al., 2024; Wang et al., 2024b), Llava v1.6 (7B, 13B, 34B) (Liu et al., 2023b,a, 2024) and QwenVL 2.5 (7B, 32B, 72B) (Bai et al., 2023; Wang et al., 2024a; Bai et al., 2025). The details of frameworks and infrastructure are given in Appendix C.

Prompt Templates: We have two decision-making prompt templates: `Simple` and `Extended`. In both prompts, we included NACE section codes and titles. In `Extended` prompt, we included the section summaries based on official guidelines with description, their content, and exclusions. Templates and contents are given in the Appendix B.

Input Configurations: In all the classification experiments, we included name of the entity in the context. In addition to this, we tested with single inputs for satellite image or external resource. As having OSM content implies at least one other resource, we did not use OSM as a single resource. We also tested combination of inputs: Satellite + OSM, Satellite + External and All.

Output Structure: We support two output structures for free text generation. In *Text* output, we instructed MLLM to generate single token answers from NACE sections or `UNK` if uncertain. To analyze the effect of classification explanations, in another output structure, we instructed MLLM to return JSON with the explanation and the decision.

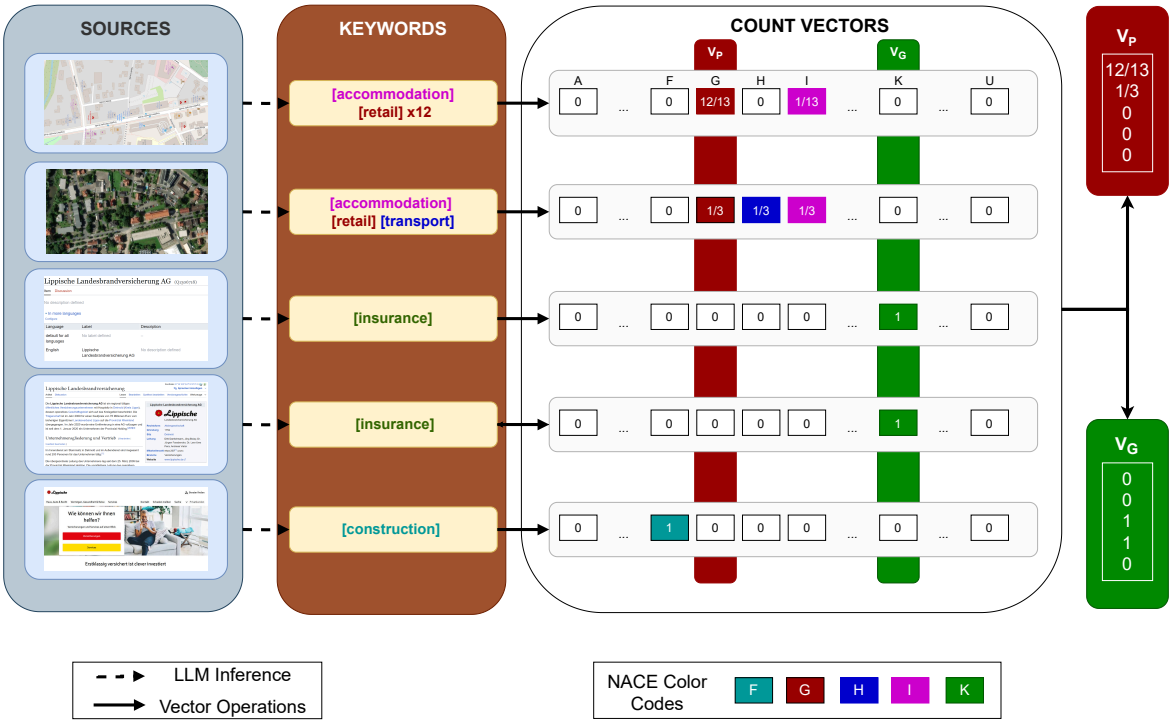

Figure 3: Example with ground truth NACE sections K and prediction G. Clues and predictions are obtained via InternVL-3 8B. Clue Analysis Methodology: (1) Sources are forwarded into separate MLLM Agents for clue extraction. (2) Keywords based on predefined Economic Activities are extracted and grouped into NACE sections. (3) For each resource (OSM, Satellite, Wikidata, Wikipedia and Website), normalized count vectors are formed. The grouped keywords are first placed to a vector with 21 dimensions (number of sections) and then divided by the total number of keywords for inference. (4) Ground-Truth and Final Prediction NACE sections are used to form V_G and V_P. Both of these vectors have the dimensions of 5 (number of inputs).

4.3 Clue Analysis

In our multi-turn pipeline, we use agents instructed to process specific input types. These agents generate free-form texts to describe economic attributes that contain keywords provided in Appendix, Table 6. For example: [retail] for section G (Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles) or [insurance] for K (Financial and Insurance Activities). In case of no evidence, agent returns No Economic Activity Found.

Through keywords, we can analyze the free-form text as shown in Figure 3. For this example, prediction is G while the correct result is K. From the satellite image, we found evidence [accommodation] (I), [retail] (G), [transport] (H). From Wikidata, we found a single keyword [insurance] (K).

By grouping these keywords based on sections, we obtain keyword counts. We scale them by the total number of keywords at inference. We refer to this scaled column vector of 21 dimensions (number of sections) as the frequency vector, $v_{i,c}$. In the case of example, the satellite frequency vector contains $1/3$ for sections G, H and I, and the Wikidata frequency vector has 1 for section K. The remaining values will be 0. If an agent fails to identify economic activity, we would have a vector of 0s.

We can use these frequency vectors to emphasize on ground truth label g, and the final prediction p. By selecting these indices, we can formulate ground truth and final prediction frequency vectors:

$$\text{Ground Truth, } v_{g_i}[i] = \begin{bmatrix} v_{i,c=\text{OSM}}[g_i] \\ v_{i,c=\text{Satellite}}[g_i] \\ v_{i,c=\text{Wikidata}}[g_i] \\ v_{i,c=\text{Wikipedia}}[g_i] \\ v_{i,c=\text{Website}}[g_i] \end{bmatrix} \quad (1)$$

$$\text{Prediction, } v_{p_i}[i] = \begin{bmatrix} v_{i,c=\text{OSM}}[p_i] \\ v_{i,c=\text{Satellite}}[p_i] \\ v_{i,c=\text{Wikidata}}[p_i] \\ v_{i,c=\text{Wikipedia}}[p_i] \\ v_{i,c=\text{Website}}[p_i] \end{bmatrix} \quad (2)$$

In the example Figure 3, we retrieve the ground truth vector via index K. For OSM, $v_{\text{OSM}}[K]$, satel-

lite $v_{\text{Satellite}}[K]$, and website, $v_{\text{Website}}[K]$, results are 0. For wikidata $v_{\text{Wikidata}}[K]$ and Wikipedia $v_{\text{Wikipedia}}[K]$ results are 1. We form the ground truth vector, $v_g = [0; 0; 1; 1; 0]$. By changing the index with prediction label, G, we can retrieve the prediction vector, $v_p = [12/13; 1/3; 0; 0; 0]$.

4.4 Metrics

In this study, in addition to *Accuracy*, we introduce *Unknown Ratio (UR)*. It is calculated using number of predictions with `UNK`, U, response over total number of inferences, I:

$$\text{Unknown Ratio (UR)} = U/I \tag{3}$$

To evaluate clue extraction, in our multi-turn pipeline, we propose additional metrics:

- **Correctness:** Measures relatedness of clues, c, to ground truth labels. It is the sum of all of ground truth vectors, $v_g[clue = c]$, divided by the number of inferences for the input, I_c.

$$\text{Correctness}_c = \frac{\sum_i^{I_c} v_{g_i}[i, c]}{I_c} \tag{4}$$

  For example, using $I_c = 1$ (due to single inference), we can find the correctness of Wikidata and Wikipedia, as $V_{G=K}[c = Wikidata] = V_{G=K}[c = Wikipedia] = 1$. Other inputs have 0 in V_G, so they have 0 correctness.

- **Effectiveness:** Measures how clues, c affect the final predictions. It is the sum of all of the prediction vectors, $v_p[clue = c]$, divided by the number of inferences for the resource, I_c.

$$\text{Effectiveness}_c = \frac{\sum_i^{I_c} v_{p_i}[i, c]}{I_c} \tag{5}$$

  In the example, only satellite and OSM have non-zero values in V_P, which are $1/3$ and $12/13$. Since we have one inference, their effectiveness will be $1/3$ and $12/13$.

5 Results

5.1 Baseline

We constructed the baseline experiments for multimodal industry classification with the Zero-Shot pipeline. `Simple` prompt template with `Text` output structure and demonstrated results for all input configurations in Table 2.

We obtained the highest performance with InternVL 3-78B model, accuracy **62.10%**, with satellite and external inputs. Among small models ($\leq 14B$), InternVL3-14B is the best-performing model. Due to its smaller context window and poorer performance, we did not use Llava v1.6 on the following experiments. The difficulty of the task is apparent as even 70B+ models failed to reach **65%** accuracy.

RQ-1: Can MLLMs use geospatial information as well as text for industry classification?

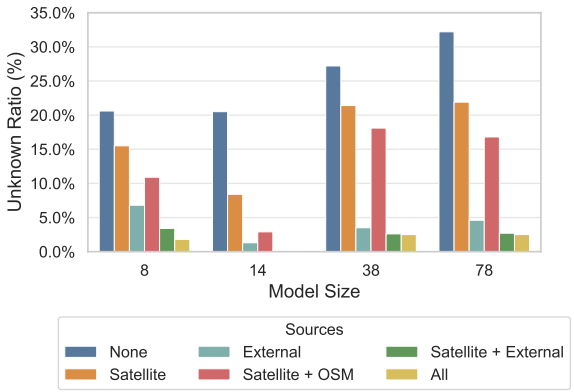

Figure 4: Baseline unknown ratio of InternVL 3 for input configurations. Unknown ratio corresponds to `UNK` responses over all inferences.

We demonstrated the capabilities of MLLM for every input configuration in Figure 4 for unknown ratio. With unknown ratio, we can measure the percentage of the responses where MLLM failed to provide NACE sections. Our two assumptions were that adding more inputs would increase performance and reduce uncertainty. However, we identified that accuracy increase is not guaranteed, especially with smaller models. Model accuracy increase does not exceed **20%** by providing inputs, which makes the name of the entity the most important resource for prediction. Furthermore, model performance is the best when external resources are given in context compared to geospatial resources.

Unlike accuracy, unknown ratio reveals that the name alone is not enough for a robust prediction. We also noted that MLLM uncertainty decreases significantly more when text information is provided compared to visual information. However, one must note that the image inputs reveal neighborhood information while the external inputs map directly to the entity.

5.2 Configurations

In our experiment setup, we allowed customization for several dimensions: output structure, prompt template, and pipeline. We selected one small ($\leq$

Model	Size (B)	None	Satellite	External	Satellite + OSM	Satellite + External	All
InternVL 2.5	1	**4.20**	2.20	1.00	2.50	0.90	0.20
	4	8.70	6.60	11.10	4.70	**13.50**	6.40
	38	46.30	49.80	58.40	51.40	**61.40**	60.10
InternVL 3	8	43.60	34.90	**48.10**	30.10	46.90	41.70
	14	45.00	49.30	**56.10**	48.30	55.60	53.00
	38	44.60	49.20	58.60	49.00	**59.80**	58.30
	78	43.40	47.80	60.40	46.10	**62.10**	58.80
Llava 1.6	7	1.50	**2.20**	✗	✗	✗	✗
	13	**13.10**	12.30	✗	✗	✗	✗
	34	1.20	**16.60**	✗	✗	✗	✗
QwenVL 2.5	7	19.80	19.10	22.10	17.60	21.80	**23.10**
	32	45.30	48.60	**57.50**	46.30	57.00	56.40
	72	46.20	43.90	56.90	45.50	59.30	**60.50**

Table 2: Baseline (*Zero-Shot* pipeline, *Simple* prompt, *Text* output) accuracy for NACE industry classification. Columns after model and size denote input configurations. Image inputs are highlighted. **Bold** indicates the best performance for the model and size pair.

$14B$) and one large ($\geq 30B$) model for InternVL 2.5 and 3 and QwenVL 2.5. The accuracy results are shown in Table 3. For these experiments, we used all the inputs available for an entry.

RQ-2: Which configuration (classification explanations, context enrichment, multi-agent) is more helpful for the task?

The smaller models perform better with the Multi-Turn pipeline. Its combination with prompt enrichment and explanations gives more than 20% boost InternVL 2.5-4B and QwenVL 2.5-7B, while InternVL 3-8B reaches almost 50% with only Multi-Turn. For larger ($\geq 30B$) models, in all the model families, we obtained the best performance with the extended prompt with classification explanation, in which no multi-agent structure is needed. The best performing smaller model is also the most recent model, InternVL 3-8B.

5.3 Clue Analysis

We extracted frequency vectors from keywords for intermediate agent *clues*. From frequency vectors, we can measure how *effective* each input is to the final prediction and how *correct* these responses are by comparing them with ground truth. For the mixture configuration (multi-turn with extended prompt and classification explanation), we calculated correctness and effectiveness. InternVL 3 (8B and 38B) results are shown in Figure 5. Other model results are available in Appendix Table 7.

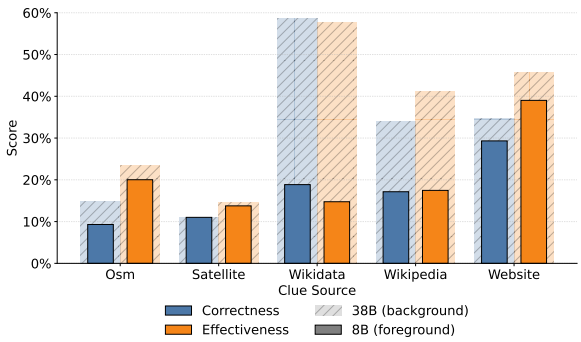

Figure 5: InternVL 3 (8B and 38B) correctness and effectiveness scores for each input. In these experiments, multi-turn pipeline with extended prompt and classification explanations is used.

RQ-3: How can we quantify intermediate agent performance with respect to the final prediction and the ground truth labels?

Both models generate more truthful clues from text sources, especially Wikidata and websites. With the exception of websites, smaller model fails to generate useful and effective clues from most sources. For the larger model, Wikidata appears to be the best resource. For image inputs, results do not improve by scale. This indicates the difficulty of clue extraction from geospatial information.

In all experiments, text input clues correlate more with the ground truth and are more effective for final prediction. This is expected because of two reasons: (1) Our text content are mostly present in or similar to the pretraining data, (2) the used models are not adapted to remote sensing.

Configuration	InternVL 2.5 4B	InternVL 2.5 38B	InternVL 3 8B	InternVL 3 38B	QwenVL 2.5 7B	QwenVL 2.5 32B
Baseline	6.40	60.10	41.70	58.30	23.10	56.40
Explanation	23.10 (16.70)	61.80 (1.70)	42.30 (0.60)	59.80 (1.50)	30.50 (7.40)	60.00 (3.60)
Extended Prompt	8.30 (1.90)	64.20 (4.10)	36.90 (-4.80)	61.30 (3.00)	27.30 (4.20)	60.40 (4.00)
Multi-Turn	22.00 (15.60)	58.20 (-1.90)	**49.80 (8.10)**	61.60 (3.30)	38.90 (15.80)	55.70 (-0.70)
Extended Prompt +	10.60 (4.20)	**65.00 (4.90)**	38.00 (-3.70)	**64.10 (5.80)**	31.00 (7.90)	**65.40 (9.00)**
Mixture	**29.20 (22.80)**	60.20 (0.10)	45.70 (4.00)	62.60 (4.30)	**45.90 (22.80)**	62.00 (5.60)

Table 3: Selected model accuracies with: Explanations, prompt context enrichment, multi-turn pipeline. Extended Prompt + is the combination of the extended prompt with explanations. Mixture is the combined setting with all the advancements. In these results, MLLMs used all inputs. The best result for a given model and size is shown in **bold**.

5.4 Ablations

Configuration	Company Websites	NAICS
Zero-Shot		
— 7B	57.93	50.19
— 32B	62.79	57.45
Few-Shot		
— 7B	58.16	51.22
— 32B	68.86	56.72
OURS LORA - 7B		
— Company Website	**89.74**	15.62
— ExioNAICS	15.76	61.44
Guo et al. (2025)		
— MiniLML3	✗	89.73
— MpnetBase	✗	**91.73**
Jagrič and Herman (2024)		
— BERT	88.23	✗

Table 4: Qwen 2.5 accuracies on text-only benchmarks: ExioNAICS (Guo et al., 2025) and Company Websites (Jagrič and Herman, 2024).

Text Only Benchmarks We validated our methodology on publicly available text-only benchmarks ExioNAICS (Guo et al., 2025) and Company Websites (Jagrič and Herman, 2024). For reproducibility, we followed their guidelines and used a fixed seed for data splitting. As they fine-tuned models, we included test results for few shots (1 sample per class) and adapted models (with LORA) in Table 4. We used Qwen 2.5 text model as it is text core for both QwenVL 2.5 and InternVL 3. Fine-tuning details are given in Appendix C.

With zero-shot pipeline, we observed similar performances for company website-based industry classification and NAICS-2 classifications. We had minimal gains with few-shot for company websites dataset and no improvement for NAICS-2. After fine-tuning our models with LORA, we surpassed Jagrič and Herman (2024) on their task.

Fine-tuning models on a specific classification scheme makes models vulnerable to future revisions. To showcase this, we tested adapted models on a given dataset on the other task and observed more than **35%** performance drop compared to zero-shot performance. Not only does fine-tuning require a significant amount of labeled data for industry classification, but fine-tuned models are also not usable after industry classification schemes change. This appears to be a major issue for the fine-grained classes. Guo et al. (2025) listed three different forms as the classification methodology changed over time. Our prompting strategy, however, is adaptable and robust for possible updates.

6 Conclusion

In this work, we introduced multimodal industry classification task with a novel dataset. Unlike its predecessors, its scope extends remote sensing benchmarks and integrates multimodal inputs to industry classification domain that was previously text-only. Our dataset, MONETA, contains entities with Satellite and OSM images, external resources and NACE sections.

We proposed a multi-turn pipeline that generates clues from each input type. Our novel metrics allowed quantitative analysis from these intermediate clues, which can assist the industry classification research moving forward. We validated our pipeline on two existing unimodal datasets and outperformed one configuration. We also showed the poorer performance of fine-tuned methods on cross-domain applications and the robustness of our zero-shot alternative. Our study indicates the difficulty of the task and textual bias of MLLMs.

Limitations

During this study, we use Gemini and ChatGPT to automatically create mappings from NACE to OSM tags. This process may introduce errors in the dataset preparation stage. In order to increase data quality, we have done extensive manual evaluation referring to the OSM wiki and NACE official guidelines.

During the data preparation of this work, NACE received another revision named as NACE Rev. 2.1. This revision split one of the major categories. Unlike prior fine-tuning approaches, our prompts can be easily modified to the new scheme and tested accordingly.

One of the limitations regarding the MLLM experiments is that some of the entities, due to initial filtering for external resources, may be in the training corpora as we include Wikidata and Wikipedia. This may be the reason behind the initially high accuracies.

We believe that future research can benefit from expert feedback and annotation in initial mapping and data quality assurance. Furthermore, the current setup can be tested with adapted MLLMs for financial and geospatial domains.

Ethics Statement

Social Impact: This work provides multimodal benchmark for industry classification task. Company entries are selected from OpenStreetMap which is publicly available. `MONETA` is intended solely for research purposes.

Dataset Access: Our code and dataset annotations are released under the Apache 2.0 and CC BY-SA 4.0 licenses, respectively. We do not hold the rights for ESRI ArcGIS World Imagery and thus will not distribute satellite images obtained from the tiles. In our datasets, we will release OSM tags and images licensed Open Data Commons Open Database License (ODbL) which also contain the bounding boxes and links to the external sources (Wikidata, Wikipedia and websites). We will also release the script to retrieve tiles and external content.

AI Assistants: AI assistants are used in this work to assist with writing by correcting grammar and code by prompt optimization and debugging.

References

Carole A Ambler and James E Kristoff. 1998. Introducing the North American industry classification system. *Government Information Quarterly*, 15(3):263–273.

Matteo Ambrois, Vincenzo Butticè, Federico Caviggioli, Giovanni Cerulli, Annalisa Croce, Antonio De Marco, Andrea Giordano, Giuliano Resce, Laura Toschi, Elisa Ughetto, and Antonio Zinilli. 2023. Using machine learning to map the european cleantech sector. EIF Working Paper Series 2023/91, European Investment Fund (EIF).

Madeline Loui Anderson, Miriam Cha, William T. Freeman, J. Taylor Perron, Nathaniel Maidel, and Kerri Cahoy. 2025. Measuring and mitigating hallucinations in vision-language dataset generation for remote sensing. In *Workshop on Preparing Good Data for Generative AI: Challenges and Approaches*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Hannah Béchara, Ran Zhang, Shuzhou Yuan, and Slava Jankin. 2022. Applying NLP Techniques to Classify Businesses by their International Standard Industrial Classification (ISIC) Code. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3472–3477.

Yuzhou Chen, Jiue-An Yang, Hugo Kyo Lee, Calvin Tribby, Tarik Benmarhnia, Marta Jankowska, and Yulia R. Gel. 2025. Fusing Multimodality of Large Language Models and Satellite Imagery via Simplicial Contrastive Learning for Latent Urban Feature Identification and Environmental Application. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Annalisa Croce, Laura Toschi, Elisa Ughetto, and Sara Zanni. 2024. Cleantech and policy framework in Europe: A machine learning approach. *Energy Policy*, 186:114006.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Guy Stephane Waffo Dzuyo, Gaël Guibon, Christophe Cerisara, and Luis Belmar-Letelier. 2025. Linking Industry Sectors and Financial Statements: A Hybrid Approach for Company Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16444–16452. Number: 16.

European Commission, editor. 2008. *NACE Rev. 2: statistical classification of economic activities in the European Community*. Publications Office, Luxembourg.

Thomas Faria and Tom Seimandi. 2023. Classifying companies in france using machine learning.

Sucharita Gopal and Josh Pitts. 2024. *Geospatial Finance: Foundations and Applications*, pages 225–273. Springer Nature Switzerland, Cham.

Yanming Guo, Xiao Qian, Kevin Credit, and Jin Ma. 2025. Group Reasoning Emission Estimation Networks. ArXiv:2502.06874 [cs].

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Timotej Jagrič and Aljaž Herman. 2024. AI Model for Industry Classification Based on Website Data. *Information*, 15(2):89. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Pu Jin, Gui-Song Xia, Fan Hu, Qikai Lu, and Liangpei Zhang. 2018. AID++: An Updated Version of AID on Scene Classification. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 4721–4724. ISSN: 2153-7003.

Sohail Ahmed Khan, Laurence Dierickx, Jan-Gunnar Furuly, Henrik Brattli Vold, Rano Tahseen, Carl-Gustav Linden, and Duc-Tien Dang-Nguyen. 2025. Debunking war information disorder: A case study in assessing the use of multimedia verification tools. *Journal of the Association for Information Science and Technology*, 76(5):752–769.

Heidi Kühnemann, Arnout van Delden, and Dick Windmeijer. 2020. Exploring a knowledge-based approach to predicting nace codes of enterprises based on web page texts. *Statistical Journal of the IAOS*, 36(3):807–821.

Haifeng Li, Hao Jiang, Xin Gu, Jian Peng, Wenbo Li, Liang Hong, and Chao Tao. 2020. CLRS: Continual Learning Benchmark for Remote Sensing Image Scene Classification. *Sensors*, 20(4):1226. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Ivan Malashin, Igor Masich, Vadim Tynchenko, Vladimir Nelyub, Aleksei Borodulin, and Andrei Gantimurov. 2024. Application of Natural Language Processing and Genetic Algorithm to Fine-Tune Hyperparameters of Classifiers for Economic Activities Analysis. *Big Data and Cognitive Computing*, 8(6):68. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Ali Mansourian and Rachid Oucheikh. 2024. ChatGeoAI: Enabling Geospatial Analysis for Public through Natural Language, with Large Language Models. *ISPRS International Journal of Geo-Information*, 13(10):348.

Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. 2024. Granular Privacy Control for Geolocation with Vision Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17240–17292, Miami, Florida, USA. Association for Computational Linguistics.

Hibiki Nakatani, Hiroki Teranishi, Shohei Higashiyama, Yuya Sawada, Hiroki Ouchi, and Taro Watanabe. 2025. A Text Embedding Model with Contrastive Example Mining for Point-of-Interest Geocoding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7279–7291, Abu Dhabi, UAE. Association for Computational Linguistics.

Sanaz Saki Norouzi and Pascal Hitzler. 2025. Knowledge-Enhanced Geospatial QA: Integrating Wikidata Fact Verification with LLMs. *Proceedings of the AAAI Symposium Series*, 5(1):334–339.

Maryan Rizinski, Andrej Jankov, Vignesh Sankaradas, Eugene Pinsky, Igor Miskovski, and Dimitar Trajanov. 2023. Company classification using zero-shot learning. ArXiv:2305.01028 [cs] version: 2.

Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. 2025. Geolocation with Real Human Gameplay Data: A Large-Scale Dataset and Human-Like Reasoning Framework. ArXiv:2502.13759 [cs].

Yuan Tao, Wanzeng Liu, Jun Chen, Jingxiang Gao, Ran Li, Xinpeng Wang, Ye Zhang, Jiaxin Ren, Shunxi Yin, Xiuli Zhu, Tingting Zhao, Xi Zhai, and Yunlu Peng. 2025. A graph-based multimodal data fusion framework for identifying urban functional zone. *International Journal of Applied Earth Observation and Geoinformation*, 136:104353.

UN. 2008. *International Standard Industrial Classification of All Economic Activities (ISIC), Rev.4*. Statistical Papers (Ser. M). United Nations, s.l.

Dimitrios Vamvourellis, Máté Tóth, Snigdha Bhagat, Dhruv Desai, Dhagash Mehta, and Stefano Pasquali. 2024. Company similarity using large language models. In *2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, pages 1–9.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Gabriela Alves Werb, Patrick Felka, Lisa Reichenbach, Susanne Walter, and Ece Yalcin-Roder. 2024. Geospatial Data and Multimodal Fact-Checking for Validating Company Data. In *2024 IEEE International Conference on Big Data (BigData)*, pages 3329–3332. ISSN: 2573-2978.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. 2017. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981.

Wenjia Xu, Zijian Yu, Yixu Wang, Jiuniu Wang, and Mugen Peng. 2024. RS-Agent: Automating Remote Sensing Tasks through Intelligent Agents. ArXiv:2406.07089 [cs].

Jeasurk Yang, Sumin Lee, Sungwon Park, Minjun Lee, and Meeyoung Cha. 2024. Poverty mapping in Mongolia with AI-based Ger detection reveals urban slums persist after the COVID-19 pandemic. ArXiv:2410.09522.

Yi Yang and Shawn Newsam. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 270–279, New York, NY, USA. Association for Computing Machinery.

A Dataset Properties

In Table 5, we listed data fields in our dataset. After filtering OSM, we identified NACE code and added category field. We extracted id, name, type and bbox from OSM fields. All the remaining attributes of OSM are present in OSM tags. After retrieving the images, we included image paths for reproducibility. We also obtained text resources (website, Wikidata or Website) and added as additional field.

Attribute	Description
`id`	Unique identifier for the object.
`type`	The object type (e.g., node, way, relation).
`name`	Human-readable name given in OSM.
`bbox`	Bounding box representing the spatial extent of the object.
`osm_tags`	OpenStreetMap tags associated with the object.
`category`	NACE Rev 2 sector classification of the entity (A to U).
`image_paths`	Dictionary of image paths for satellite and OSM images.
`sources`	Dictionary of external sources (Website Text, Wikipedia Text, Wikidata JSON).

Table 5: Dataset entry contents. Each entry of `MONETA` contains these attributes. Id, type, name, bbox and OSM tags are retrieved from OSM. Sources are retrieved online from existing OSM tags.

A.1 NACE to OSM Mapping

In this part, we provide an example for the data preparation process given in Figure 2. First, we retrieve the title, content, scope, additional content and exclusion from RDF/XML as shown in Figure 8. We use these contents with gemini to generate OSM tags list, Figure 9. We validated generated lists by searching the entries in OSM tags database and cross-checked with NACE guidelines. In case we find other relevant tags, we added those instances to the list. For the example of Finance and Insurance Activities we obtained following tags:

- 'amenity=insurance'
- 'office=insurance'
- 'amenity=bureau_de_change'
- 'office=financial'
- 'office=accountant'

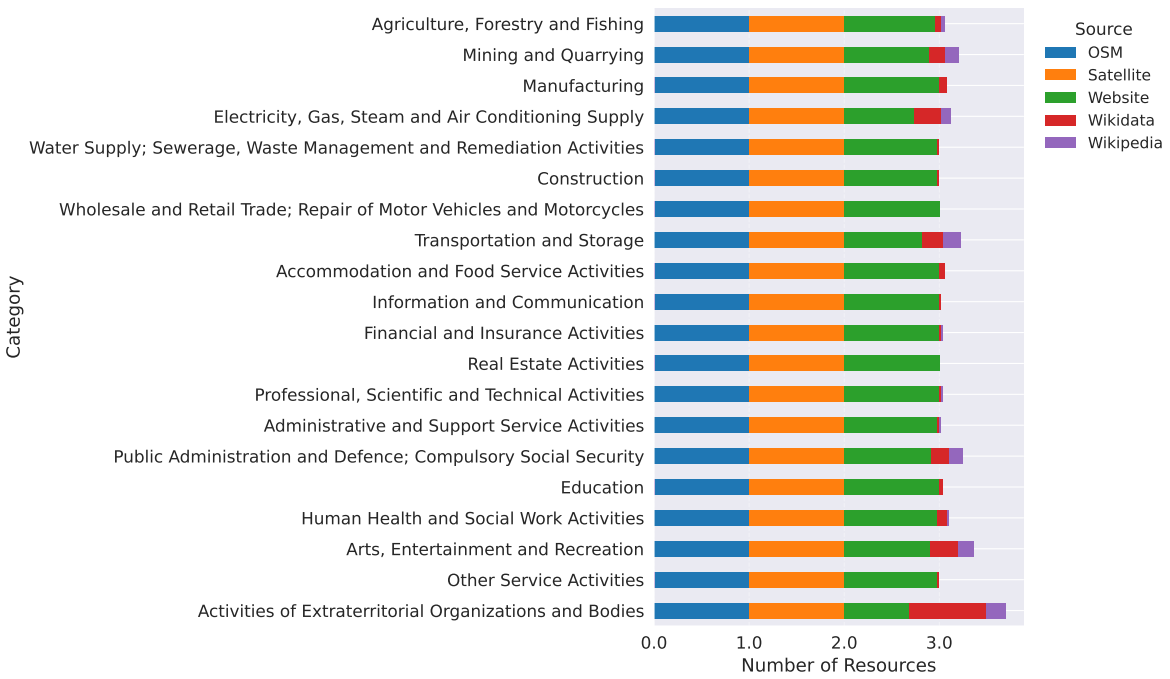

Figure 6: MONETA mean resource per category

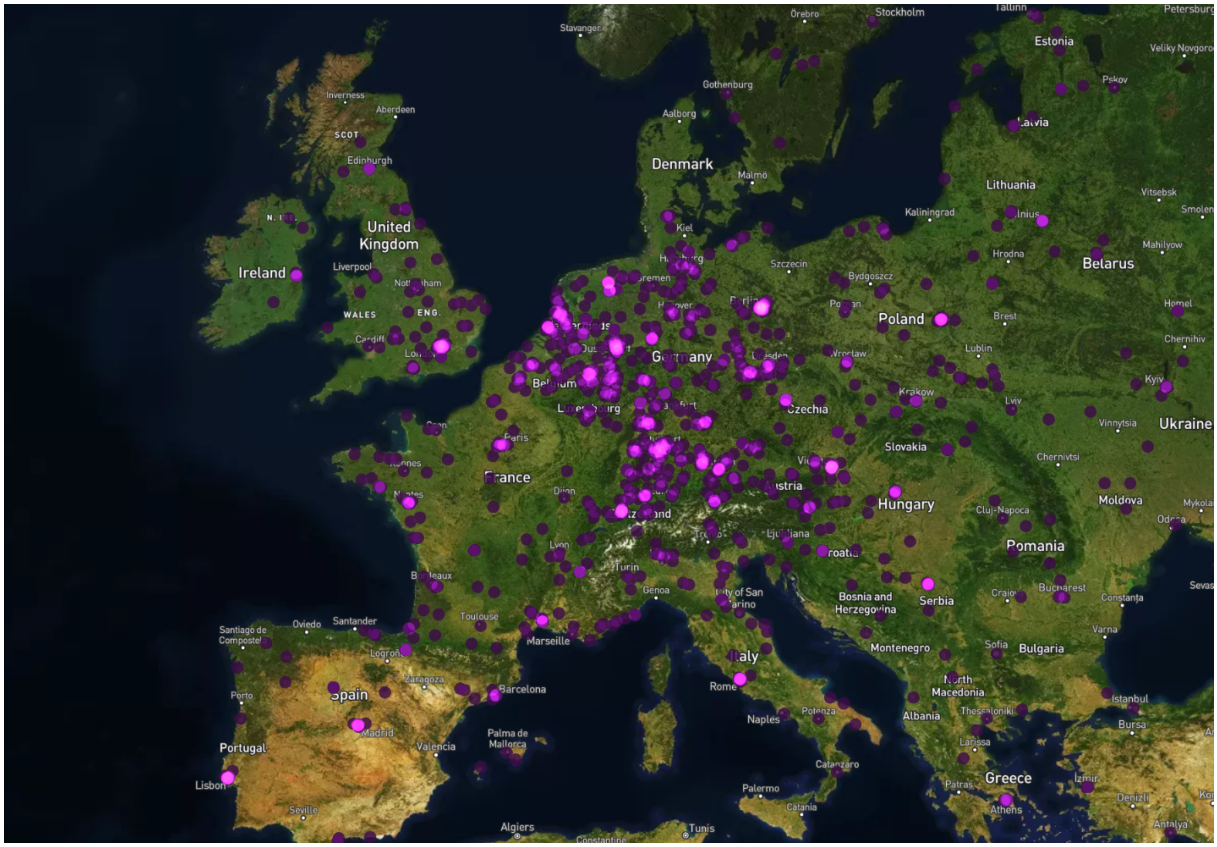

Figure 7: MONETA entries around Europe

- 'office=tax_advisor'
- 'shop=currency_exchange'

We defined custom tags for data quality for address, external data pointers and name of the entity in addition to these attribute. For address tags, we

used house number, street and postal code keywords as the other attributes can be obtained via reverse geocoding. For each NACE code, we query European OSM extract with corresponding OSM tags list.

A.2 NACE Details

We extracted NACE codes, titles, descriptions and keywords for prompting. Summary of these attributes are given in Table 6.

B Prompts

B.1 Data Preparation Prompts

We use Gemini to create OSM tags list from NACE RDF/XML descriptions. This results are manually checked to create OSM Tags lists for each NACE section.

B.2 Output Prompts

We have two output prompts. The `Text` prompts instructs MLLM to return a single token answer. It is used in the baseline configuration. If an MLLM cannot identify the class it returns the class as `UNK`. The other configuration, used for explanations, is `JSON` output prompt. The JSON output starts with the explanation less than 50 words and followed by the classification decision.

B.3 Zero Shot

In the Zero-Shot classification prompt, we define instructions for the input types to guide MLLMs feature extraction process. Then we provide NACE_CONTEXT. This context can be `Simple` (NACE Codes and Titles) and `Extended` (NACE Codes, Titles and AI generated summaries from official guidelines).

In the classification prompt, we define the set of available inputs and as we test the models without any inputs, we instruct model to use name if no input is provided. Then, we provide the output prompt depending the experiment setup. Finally, we give the context based on input configuration.

B.4 Multi Turn

In our muli-turn pipeline, we have intermediate level agents for each input type (Satellite, OSM, Wikidata, Wikipedia, and Website). Each processor agent prompt, contains several instructions for data processing and sets the generation limit to 512 tokens. MLLMs are instructed to generate **No Economic Activity Found** in response, if they cannot retrieve evidence from an input. For each processor, we give NACE Keywords defined in Table 6. These keywords are expected in the output for easier grouping of the free form text.

Clues are appended to the Multi-Turn classification prompt after construction. Final decision making agent prompts resembles the single stage Zero-Shot classification prompt. It replaces the entries with the text clues.

B.5 Ablation Prompts

Based on our zero-shot template, we designed prompts for text only benchmarks ExioNAICS (Guo et al., 2025) and Company Websites (Rizinski et al., 2023). For the few-shot examples, we selected randomly selected with fixed seed one example per class from the training set.

C Implementation Details

C.1 Frameworks

In order to run models, we preferred Hugginface's Transformers, (Wolf et al., 2020) library due to multimodal support. During the ablation studies on text only benchmarks, we used Unlsoth (Daniel Han and team, 2023) to train models with LoRA (Hu et al., 2022).

C.2 Infrastructure

In our experiments, we used NVIDIA A100 40GB GPUs and increased number of GPUs depending on the model size.

C.3 Hyperparameters

During the ablation studies for the text only benchmarks, we fine tuned models with LoRA (Hu et al., 2022) using rank 32 and alpha 64 for 5 epochs with learning rate $2e-4$.

D Additional Results

D.1 Baseline Results

The model performances across the models increase significantly until 30B+ variants. However, the performance increase between the largest counterparts is negligible in most cases compared to the overall trend. We demonstrate this effect of model size on the performance in Figure 10.

D.2 Section-wise Analysis

Based on the available configurations (Baseline, Explanation, Extended Prompt, Multi-Turn, Extended Prompt + (Explanation), Mixture), we cre-

NACE RDF/XML Extract for K

XML/RDF content at https://publications.europa.eu/resource/authority/ux2/nace2/K is retrieved as such:

```
{
    'Official Name': 'K FINANCIAL AND INSURANCE ACTIVITIES',
    'Alternative Name': 'FINANCIAL AND INSURANCE ACTIVITIES',
    'Scope': None,
    'Content': 'This section includes financial service activities, including
    insurance, reinsurance and pension funding activities and
    activities to support financial services.',
    'Additional Content': 'This section also includes the activities of holding
    assets, such as activities of holding companies and
    the activities of trusts, funds and similar financial entities.',
    'Exclusion': None
}
```

Figure 8: NACE XML/RDF extract for Financial and Insurance Activities

Gemini OSM Tags for K

"amenity=bank", "amenity=atm", "amenity=bureau_de_change", "shop=money_lender", "shop=insurance", "office=financial", "office=insurance", "office=financial_advisor", "office=company", "company=insurance", "office=consulting"

Figure 9: Gemini generated OSM tag list for Financial and Insurance Activities. Relevant tags are highlighted. Other tags either do not match with official guidelines, rare or non-existing.

Data Preparation prompt for NACE to OSM Mapping

You will be given a description for a NACE code, which describes business activities. Your job is to identify relevant OSM (OpenStreetMap) tags that could be used to classify businesses or locations related to this NACE code. You will provide the tags in a Python list format.
NACE Code Description:
{RDF/XML Extract}
Your response should ONLY be a Python list of OSM tags. Do not include any other text or explanation. The list should contain strings. For example:
["landuse=retail", "shop=supermarket", "amenity=parking"]
Instructions:
- DO NOT include relations like this landuse, shop, amenity but tags like this landuse=retail, shop=supermarket, amenity=parking. All tags will have = sign.
- DO NOT PUT ANY EXPLANATION IN THE PROMPT.
- DO NOT PUT PYTHON INDICATOR IN THE PROMPT.
- DO NOT USE ANY TAGS THAT DOES NOT DESCRIBE A BUSINESS ACTIVITY LIKE LANDUSE=FOREST. - ONLY PROVIDE THE LIST OF TAGS IN PYTHON LIST FORMAT.
OSM TAGS=

ated confusion matrices between ground truth and prediction results. We retrieved the counts from the diagonals and visualized it with respect to configurations in Figure 11. We observed poorer per-

Section	Title	Description	Keywords
A	Agriculture, Forestry and Fishing	This section covers the utilization of plant and animal natural resources through farming, animal husbandry, and harvesting from natural environments.	[agriculture], [forestry], [fishing], [crops], [livestock], [timber]
B	Mining and Quarrying	This section includes the extraction of naturally occurring minerals in solid, liquid, or gaseous forms, using various methods such as underground mining, surface mining, and well operations, along with related preparation activities.	[mining], [quarrying], [oil], [coal], [ores]
C	Manufacturing	This section includes the physical or chemical transformation of raw materials or components into new products, typically resulting in outputs ready for use or as inputs to further manufacturing.	[manufacturing], [processing], [assembly], [fabrication]
D	Electricity, Gas, Steam and Air Conditioning Supply	This section covers the provision and distribution of electricity, natural gas, steam, hot water, and air conditioning through a permanent infrastructure of networks such as lines, mains, and pipes.	[electricity], [gas], [steam]
E	Water Supply; Sewerage, Waste Management and Remediation Activities	This section includes the collection, treatment, and disposal of waste and sewage, as well as the management of contaminated sites and the supply of water for various uses.	[water], [sewerage], [waste], [remediation]
F	Construction	This section covers general and specialised construction activities for buildings and civil engineering works, including new projects, repairs, additions, and temporary structures, whether performed directly or through subcontracting.	[construction], [building], [infrastructure]
G	Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles	This section includes the wholesale and retail sale of goods without transformation and related services, as well as the repair of motor vehicles and motorcycles.	[wholesale], [retail], [trade], [resale], [vehicle-repair]
H	Transportation and Storage	This section includes the transport of passengers or freight by various modes, along with related services such as cargo handling, storage, and postal and courier activities.	[transport], [logistics], [freight], [storage], [postal]
I	Accommodation and Food Service Activities	This section covers short-term accommodation services for travelers and the preparation and serving of meals and drinks for immediate consumption.	[accommodation], [hotels], [restaurants], [catering]
J	Information and Communication	This section includes the creation, publishing, and distribution of information and cultural content, telecommunications, IT services, and data processing activities.	[information], [communication], [telecom], [publishing], [IT]
K	Financial and Insurance Activities	This section includes activities related to financial services, insurance and pension funding, and asset-holding entities such as holding companies and trusts.	[finance], [insurance], [banking], [investment]
L	Real Estate Activities	This section includes activities related to real estate sales, rentals, management, and related services, carried out either on owned or leased property or on a contract basis.	[real-estate], [property], [leasing]
M	Professional, Scientific and Technical Activities	This section includes specialised services requiring high levels of expertise, such as legal, accounting, engineering, and scientific research services.	[professional], [scientific], [technical], [legal], [engineering], [research]
N	Administrative and Support Service Activities	This section includes support services for general business operations that do not primarily involve the transfer of specialised knowledge, such as employment services, security, and facility management.	[administration], [support], [employment], [security], [cleaning]
O	Public Administration and Defence; Compulsory Social Security	This section includes government-related activities such as legislation, taxation, national defence, public order, immigration, foreign affairs, and compulsory social security administration.	[government], [defence], [legislation], [taxation]
P	Education	This section includes all levels and types of education, from preschool to higher education, including adult and special education, whether provided publicly or privately, through various formats such as in-person or online.	[education], [training], [schooling]
Q	Human Health and Social Work Activities	This section includes medical care by health professionals, residential care involving health support, and social work activities without health care involvement.	[health], [social-care], [medical], [hospitals], [clinics]
R	Arts, Entertainment and Recreation	This section includes cultural, artistic, entertainment, and recreational activities for the general public, including live shows, museums, gambling, sports, and leisure facilities.	[arts], [entertainment], [recreation], [sports], [culture]
S	Other Service Activities	This section includes a variety of personal services not classified elsewhere, such as those provided by membership organisations and the repair of computers and household goods.	[personal-services], [household-services], [memberships], [repairs]
T	Activities of Households as Employers; Undifferentiated Goods- and Services-producing Activities of Households for Own Use	This section includes households' subsistence production of goods and services for their own use, when no primary activity can be identified and the output is not for market sale.	[household-employment], [household-production]
U	Activities of Extraterritorial Organisations and Bodies	This section includes the activities of international organisations such as the UN, IMF, World Bank, and diplomatic missions determined by the host country location.	[extraterritorial], [embassies], [diplomacy]

Table 6: NACE Codes, Titles, AI Generated descriptions and keywords (from official guidelines)

Text Output Prompt

You will return only the SECTOR CODE. If you are not sure about the sector code, return "UNK" as a default value.
For example: A
SINGLE TOKEN RESPONSE ONLY

JSON Output Prompt

You will return a JSON output including the Sector and Explanation. Explanation should be a short description, less than 50 words, of why you chose this sector code.

```
{
    "EXPLANATION": "This belongs to Category A because ...",
    "LLM_RESPONSE": "A"
}
```

DO NOT PRINT ANYTHING OTHER THAN JSON RESPONSE

Zero Shot Classification Prompt

You are a helpful assistant that designed to understand economical activities from various resources. These resources can be OpenStreetMap (OSM) and Satellite Imagery, Sources online such as Wikidata, Wikipedia and/or Website.
From the images, you will to analyze and identify geospatial features such as but not limited to:
- Buildings
- Terrains
- Streets

From the sources, you will analyze and identify the context related to economical activity such as:
- Products
- Services
- Activities
- Business Type
- Industry

Based on these attributes and the given name of the entity you will identify NACE Economical Activity sector code.
{NACE_CONTEXT}
Following sources will be provided: {resources}
Note that you may not be given all of the sources.
osm: OSM Image
satellite: Satellite Image
source: Source (Wikidata, Wikipedia, Website)

If no sources are provided, you will use the name of the entity to identify NACE Economical Activity sector code.
{OUTPUT_FORMAT}

formances for the sections after M in the smaller models, and 30B+ models are overall consistent with the exceptions of sections F (Construction), N (Administrative and Support Service Activities) and S (Other Service Activities).

D.3 Experiment Dimension Analysis

We visualized the effect of experiment ensembles in our Figure 12. In the x-axis, we incrementally added the results. Therefore, it corresponds to Baseline, Explanation, Extended Prompt + Explanation, and Multi-Turn + Extended Prompt + Explanation. In addition to accuracy, we used precision and recall. In all these metrics, changes in smaller models are superior compared to changes in larger models in the same family.

OSM Image Processor Agent Prompt

You are analyzing an OpenStreetMap (OSM) map description to identify explicit mentions of economic activities.
Instructions:
- Identify labeled features, POIs, or tags that clearly indicate economic activity.
- The keyword list below represents valid economic activity categories. Only mention an activity if there is direct evidence in the OSM data.
- Match terms in the description only if they clearly correspond to one of the keywords below or its direct synonym.
- When you refer to an activity, wrap it in [ ] exactly as shown in the keyword list.
- For each activity found, also mention the specific OSM feature name, tag, or label that supports it.
- Do not invent, infer, or guess activities not explicitly visible or stated.
- Keep the total output under 512 tokens.
- If no economic activity is present, output exactly: "No economic activity clues found."
- Output language must be English

{ECONOMIC ACTIVITY KEYWORDS}

Output format:
Economic activity clues:
- [keyword] — OSM feature/tag that supports it
OSM map description:

Satellite Image Processor Agent Prompt

You are analyzing a satellite image description to identify explicit mentions of economic activities.
Instructions:
- Identify labeled features, POIs, or icons that clearly indicate economic activity.
- The keyword list below represents valid economic activity categories. Only mention an activity if there is direct evidence in the description.
- Match terms in the description only if they clearly correspond to one of the keywords below or its direct synonym.
- When you refer to an activity, wrap it in [ ] exactly as shown in the keyword list.
- For each activity found, also mention the specific labeled feature, POI name, or visible element that supports it.
- Do not invent, infer, or guess activities not explicitly visible or stated.
- Keep the total output under 512 tokens.
- If no economic activity is present, output exactly: "No economic activity clues found."
- Output language must be English

{ECONOMIC ACTIVITY KEYWORDS}

Output format:
Economic activity clues:
- [keyword] — labeled feature/POI from satellite image that supports it
Satellite image description:

Website Processor Agent Prompt

You are analyzing website content to identify explicit mentions of economic activities.
Instructions:
- Only consider explicitly stated fields such as "industry", "product", "field of work", or equivalent.
- The keyword list below represents valid economic activity categories. Only mention an activity if there is direct evidence in the website content.
- Match terms in the content only if they clearly correspond to one of the keywords below or its direct synonym.
- When you refer to an activity, wrap it in [ ] exactly as shown in the keyword list.
- For each activity found, also mention the specific entity, phrase, or value from the website that supports it.
- Do not invent, infer, or guess activities not explicitly stated.
- Keep the total output under 512 tokens.
- If no economic activity is present, output exactly: "No economic activity clues found."
- Output language must be English

{ECONOMIC ACTIVITY KEYWORDS}

Output format:
Economic activity clues:
- [keyword] — entity/phrase from website that supports it
Website content:

Wikidata Processor Agent Prompt

You are analyzing a Wikidata JSON entry to identify explicit mentions of economic activities.
Instructions:
- Only consider explicitly stated fields such as "industry", "product", "field of work", or equivalent.
- The keyword list below represents valid economic activity categories. Only mention an activity if there is direct evidence in the data.
- Match terms in the entry only if they clearly correspond to one of the keywords below or its direct synonym.
- When you refer to an activity, wrap it in [ ] exactly as shown in the keyword list.
- For each activity found, also mention the specific entity or value from the Wikidata entry that supports it.
- Do not invent, infer, or guess activities not explicitly stated.
- Keep the total output under 512 tokens.
- If no economic activity is present, output exactly: "No economic activity clues found."
- Output language must be English

{ECONOMIC ACTIVITY KEYWORDS}

Output format:
Economic activity clues:
- [keyword] — entity/value from Wikidata that supports it
JSON:

D.4 Clue Ablations

Model section preferences From clue frequency vectors, we identified percentages for each NACE

Wikipedia Processor Agent Prompt

You are analyzing a Wikipedia article to identify explicit mentions of economic activities.
Instructions:
- Only consider explicitly stated fields such as "industry", "product", "field of work", or equivalent.
- The keyword list below represents valid economic activity categories. Only mention an activity if there is direct evidence in the article.
- Match terms in the article only if they clearly correspond to one of the keywords below or its direct synonym.
- When you refer to an activity, wrap it in [ ] exactly as shown in the keyword list.
- For each activity found, also mention the specific entity, phrase, or value from the article that supports it.
- Do not invent, infer, or guess activities not explicitly stated.
- Keep the total output under 512 tokens.
- If no economic activity is present, output exactly: "No economic activity clues found."
- Output language must be English

{ECONOMIC ACTIVITY KEYWORDS}

Output format:
Economic activity clues:
- [keyword] — entity/phrase from Wikipedia that supports it
Wikipedia content:

Multi-Turn Classification Prompt

You are a helpful assistant that designed to understand economical activities from various resources. You will be provided with clues from different sources such as Wikidata, Wikipedia, websites, OpenStreetMap (OSM) images, and satellite images. Based on these attributes and the given name of the entity you will identify NACE Economical Activity sector code.
{NACE_CONTEXT}
Note that you may not be given all of the clues. And if you are not given any clues, you should identify based on the name of the entity.
{OUTPUT_FORMAT}

sections for InternVL 3 (8 and 38B) and QwenVL 2.5 (7B and 32B). Regardless of the architecture and model size we observe a strong representation of *Wholesale and Retail Trade*, *Transportation and Storage* and *Accommodation and Food Service Activities*. In the original dataset, the distribution of visual sources are uniform. However for clues, bias is observed for the listed categories. In addition to this, as it was shown in Figure 6, Wikidata and Wikipedia were dominant sources in the last category. While, smaller models fail to utilize these sources, larger models clearly extract correct clues. We visualized this in Appendix as a confusion matrix, In Figure 13.

D.5 Information Discovery

We instructed clue extractors to return No Economic Activity Found in case there is no evidence in the source. Using the number of inferences with this phrase, NEI_c, we defined **Information Discovery Rate** as:

$$\text{Information Discovery } (ID_c) = 1 - \frac{NEI_c}{I_c} \quad (6)$$

I_c denotes the total number of inferences per clue, c. The metric demonstrates the evidence retrieved from a input type c. It is scaled between 0 and 1.
Information discovery from clues In Table 8, Information Discovery Rates are given. Smaller mod-

ExioNAICS Classification Prompt

INSTRUCTIONS: Classify the company into one industry sector.
You are given codes and titles.
Respond with EXACTLY ONE UPPERCASE LETTER from A–T. Do NOT include spaces, newlines, punctuation, or any other text.
If unsure, pick the single best letter based on the company's **primary revenue-generating activity**.
VALID LETTERS: A B C D E F G H I J K L M N O P Q R S T

Choices (A–T): Title
A: Agriculture, Forestry, Fishing and Hunting
B: Mining, Quarrying, and Oil and Gas Extraction
C: Utilities
D: Construction
E: Manufacturing
F: Wholesale Trade
G: Retail Trade
H: Transportation and Warehousing
I: Information
J: Finance and Insurance
K: Real Estate and Rental and Leasing
L: Professional, Scientific, and Technical Services
M: Management of Companies and Enterprises
N: Administrative and Support and Waste Management and Remediation Services
O: Educational Services
P: Health Care and Social Assistance
Q: Arts, Entertainment, and Recreation
R: Accommodation and Food Services
S: Other Services (except Public Administration)
T: Public Administration

Model	Size (B)	OSM		Satellite		Wikidata		Wikipedia		Website	
InternVL 2.5	4	1.74	2.62	6.26	**8.46**	6.56	0.82	**14.54**	7.85	7.61	6.47
	38	8.15	16.05	6.58	8.12	**54.10**	**56.28**	37.58	38.83	30.03	43.45
InternVL3	8	9.31	20.02	11.02	13.77	18.85	14.75	17.15	17.47	**29.32**	**39.03**
	38	14.77	23.40	11.04	14.62	**58.61**	**57.65**	33.92	41.23	34.58	45.72
Qwen VL 2.5	7	12.81	21.49	10.30	15.85	5.74	3.28	19.23	15.38	**19.98**	**25.10**
	32	18.13	26.68	10.36	13.15	**54.37**	**57.65**	22.34	24.05	32.40	42.84

Table 7: Correctness (left) and Effectiveness (right) by inputs for Mixture. Image inputs are highlighted.

els of InternVL 2.5 and QwenVL 2.5 may fail to extract information. The information discovery highlights architectural differences. InternVL 2.5 and 3 discovered more information from texts more while QwenVL 2.5 discovered most of the information from OSM. According to the results, InternVL 3 has a stronger vision core of compared to its predecessor, as it is now able to utilize satellite and OSM images more than **30%** at the same size.

Company Websites Classification Prompt

INSTRUCTIONS: Classify into one category.
Respond with EXACTLY ONE UPPERCASE LETTER from A–M. Do NOT include spaces, newlines, punctuation, or any other text.
If unsure, pick the single best letter based on the company's **primary revenue-generating activity**.

VALID LETTERS: A, B, C, D, E, F, G, H, I, J, K, L, M
Choices (A–M): Title:
A: Commercial Services & Supplies
B: Healthcare
C: Materials
D: Financials
E: Energy & Utilities
F: Professional Services
G: Corporate Services
H: Media, Marketing & Sales
I: Information Technology
J: Consumer Discretionary
K: Industrials
L: Transportation & Logistics
M: Consumer Staples

Model	Size (B)	OSM	Satellite	Wikidata	Wikipedia	Website
InternVL 2.5	4	24.50	35.20	13.11	**51.92**	22.85
	38	59.20	20.70	71.31	**84.62**	77.47
InternVL 3	8	84.60	77.50	31.15	67.31	**91.82**
	38	87.10	78.40	78.69	**90.38**	79.38
QwenVL 2.5	7	**64.70**	47.10	11.48	44.23	60.26
	32	**89.90**	84.90	75.41	76.92	78.75

Table 8: Information Discovery of inputs. Image inputs are highlighted. **Bold** denotes the highest information discovery from a model-size pair.

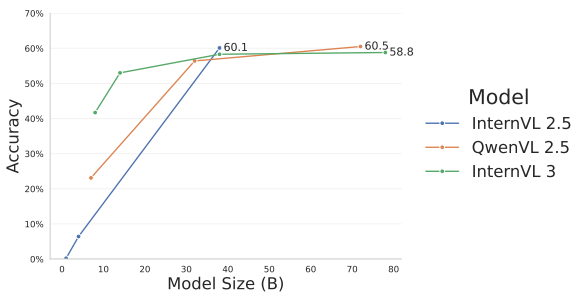

Figure 10: Baseline Accuracies by Model Size for ALL input Configuration

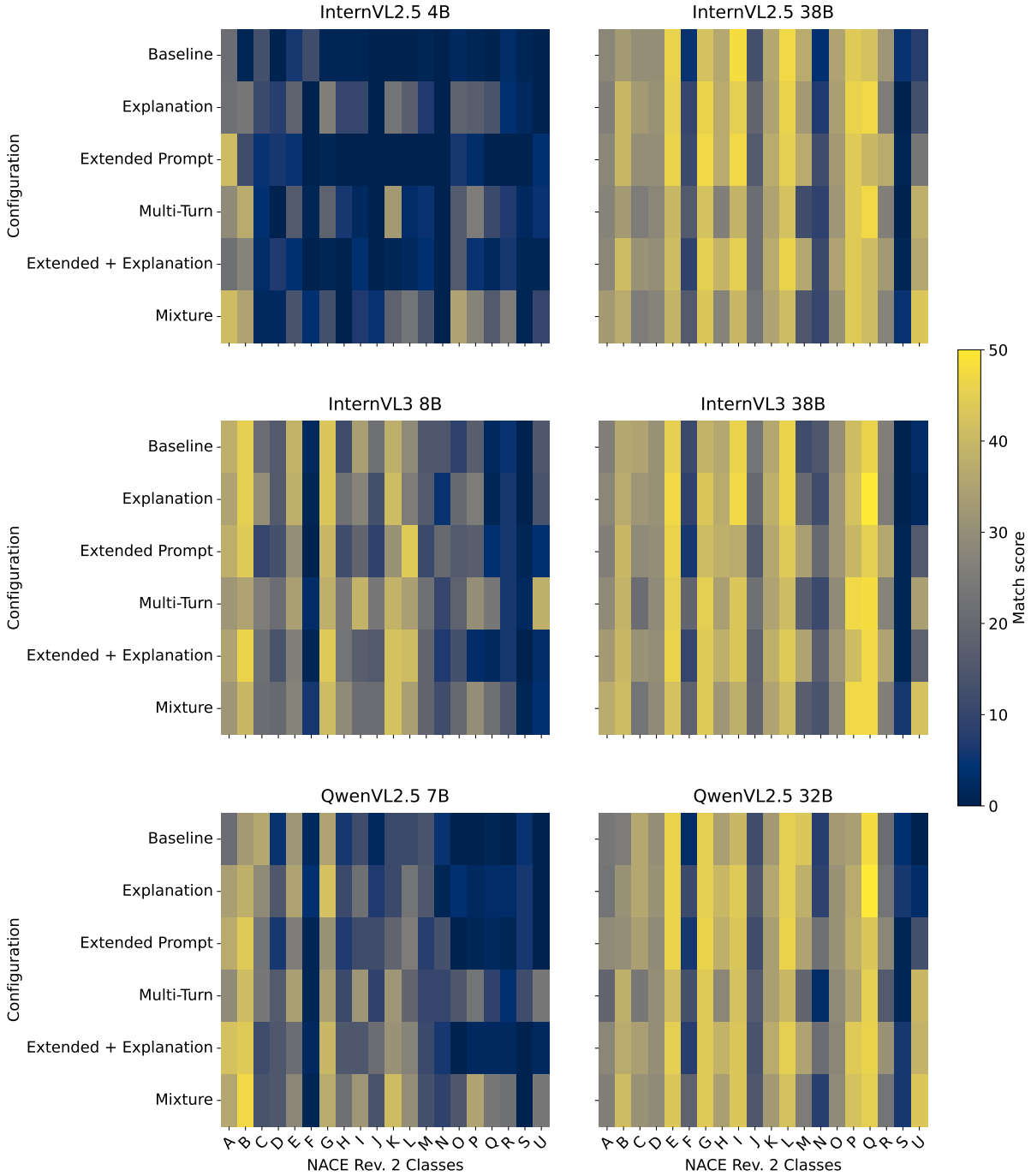

Figure 11: NACE Section-wise ground truth-prediction matches for available experiment dimensions. Mixture denotes the combination of all improvement methodologies.

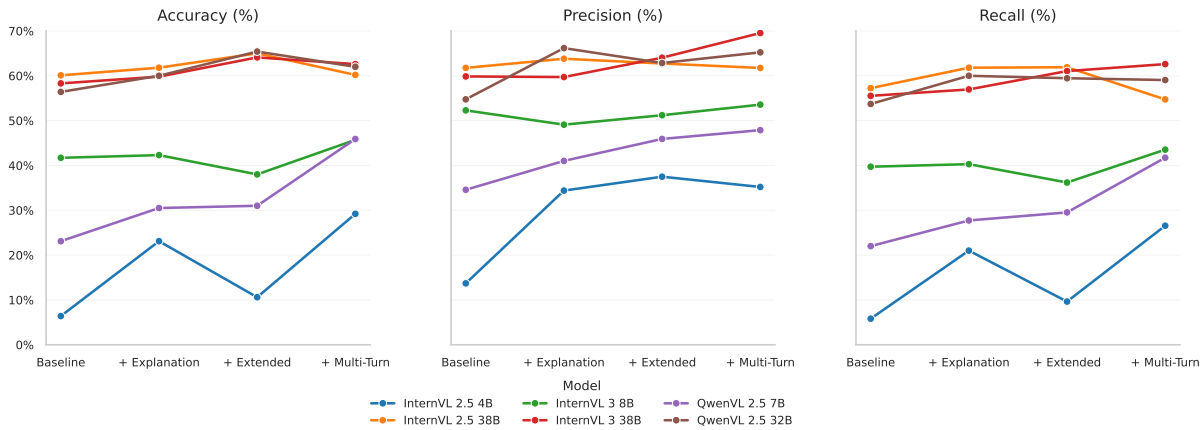

Figure 12: Model Performance changes in Accuracy, Precision and Recall

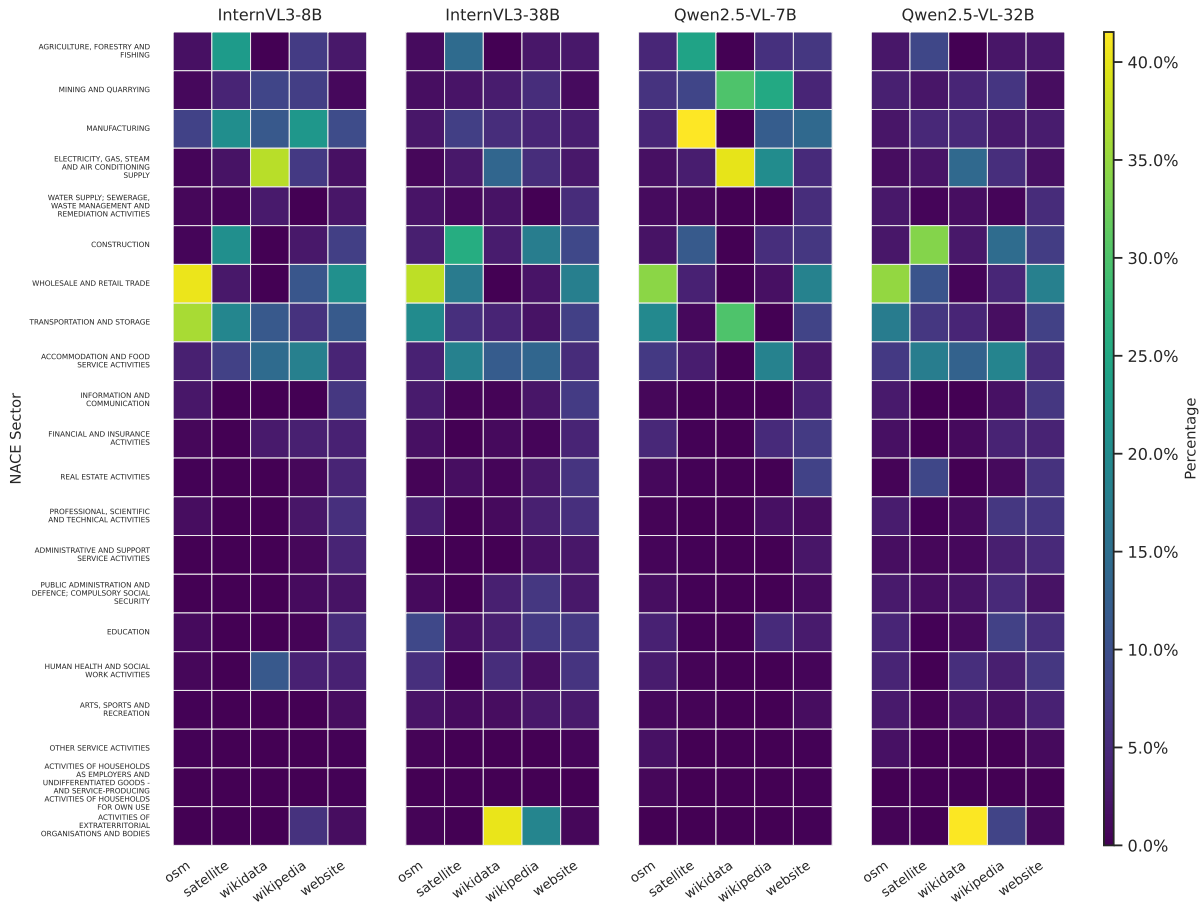

Figure 13: Clue Keyword NACE sections for selected Models