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 with text (Website, Wikipedia, Wikidata) and geospatial sources (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 baseline reaches 62.10% and 74.10% with open and closed-source Multimodal Large Language Models (MLLM). We observe 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. One of the important milestones in multimodal AI research, Multimodal Large Language Models (MLLM) such as Llava (Liu et al., 2023b,a, 2024), InternVL (Chen et al., 2024; Wang et al., 2024b), QwenVL (Bai et al., 2023; Wang et al., 2024a; Bai et al., 2025), GPT-5 (OpenAI, 2025) and Gemini 2.5 (Gemini 2.5 Team, 2025) can contribute to geospatial finance decision making tasks by processing visual geospatial data in addition to text documents. Traditionally, AI research developed unimodal automatic industry classification, a research 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.

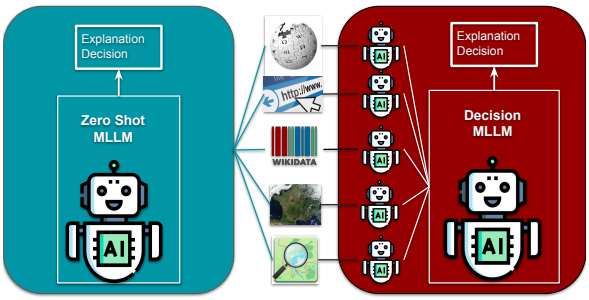


Figure 1: Zero-Shot vs Multi-Turn comparison for MONETA. (1) Zero-shot pipeline (left): Available resources are forwarded into the industry classifier MLLM together. Explanations and classifications are obtained. (2) Multi-Turn pipeline (right): Each resource is processed by separate specialized agents.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

Due to the ever-changing nature of businesses with time and location, several industry classification 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 methods rely solely on text, which is often unavailable for newly founded or small firms, whereas geospatial information from business registers can provide useful signals. Second, they fine-tune models that require large datasets and limit them to a single classification scheme.

We connect economic activities to spatial extent

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.

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?

In this study, we propose a novel task: **Multimodal Industry Classification with Geospatial Information** and introduce:

- **MONETA: Multimodal Industry Classification Benchmark** for 1,000 European businesses in 20 NACE (European Commission, 2008) sections. We provide two visual resources (OpenStreetMap (OSM) and Satellite) and at least one text resource (Wikidata, Wikipedia, and website) per entry.
- **Multimodal Industry Classification**: An expert domain multimodal AI task rooted in geospatial finance. We propose Zero Shot and Multi-Turn (Multi-Agent) approaches supporting various multimodal resources, 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 dates back to the 1930s with the Standard Industrial Classification (SIC), supporting market and sustainability analysis (Ambrois et al., 2023; Croce et al., 2024). Automated business classification remains an active research area (Kühnemann et al., 2020; Faria and Seimandi, 2023). To reflect emerging sectors, multiple schemas have been introduced, including GICS by MSCI and S&P,[1] and ISIC by the United Nations (UN, 2008), with regional variants such as NAICS (Ambler and Kristoff, 1998) and NACE (European Commission, 2008). The European Union uses NACE, a hierarchical ISIC-based scheme with 21 sections (A–U) representing major economic activities (e.g., *C: Manufacturing*), followed by 88 divisions, 272 groups, and 514 classes.

For automatic industry classification, Kühnemann et al. (2020) used websites for NACE classes. Rizinski et al. (2023), Faria and Seimandi (2023) and Vamvourellis et al. (2024) used company descriptions to classify industries. Fine-tuning transformers and adapters is a common approach (Béchara et al., 2022; Jagrič and Herman, 2024; Guo et al., 2025; Dzuyo et al., 2025) for coarse and fine-grained industry classification. 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 and makes the model unusable for other schemas. Furthermore,

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

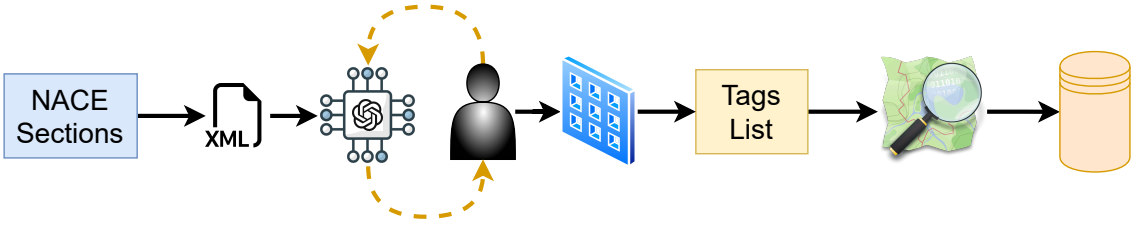


Figure 2: Overview of the dataset preparation process. (1) NACE section XMLs are initially converted to the OSM tags and manually checked by the authors. (2) We added custom filters for data quality and queried Europe OSM data with tag list. (3) Samples are grouped by NACE codes to form the gold dataset.

all the studies incorporated unimodal text sources, such as financial statements, which may not be available 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), Question 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).

Remote sensing extracts geospatial features from sources such as satellite imagery, street views, and OpenStreetMap (OSM[2]), which can link to external resources like Wikipedia, Wikidata, and websites. The AI community has developed many remote sensing datasets, including UC-MED (Yang and Newsam, 2010) for land-use classification, AID and AID++ (Xia et al., 2017; Jin et al., 2018) for aerial scene understanding, and CLRS (Li et al., 2020) for continual learning.

Recent GeoAI work fuses multimodal data to infer economic attributes in urban contexts (Tao et al., 2025; Chen et al., 2025). For example, Yang et al. (2024) linked geospatial data to economic activity for poverty mapping, and Li et al. (2025) combined spatial and temporal signals for health and public-service traffic accident analysis.

Despite these advances, prior work does not address entity-level industry classification, as reflected in Table 1. Our work is the first to study the suitability of geospatial resources for industry classification of individual businesses.

3 MONETA

We introduce MONETA, a novel multimodal benchmark for industry classification based on EU Guidelines (NACE) for European businesses. In this section, we explain our mapping and dataset.

Mapping: Due to a lack of direct mapping connecting OSM and NACE sections, we generated a **novel NACE to OSM** mapping using the methodology shown in Figure 2.

We first used Gemini to generate OSM tags for NACE sections from official guidelines in RDF/XML. Because this mapping was error-prone, we introduced a human-in-the-loop process and refined the annotations using GPT and Gemini. This resulted in a validated list of OSM tags per NACE section. We then applied data-quality filters (name, address, and external links such as website, Wikipedia, and Wikidata). Finally, we queried OSM, grouped entries by NACE section, and obtained the gold dataset. Additional details and examples are given in Appendix A.1.

Gold Dataset: We sampled 50 entries per NACE section (A–U, excluding T) and formed the first multimodal industry classification benchmark with 1,000 businesses. Using bounding boxes, we computed dynamic zoom levels and retrieved satellite imagery via the ESRI REST API.[3] ESRI and OSM services return static tiles for given coordinates and zoom levels; concatenating these tiles yields aligned OSM and satellite images of the same area. None of the external resources in MONETA explicitly mentions the NACE section in their context. Additional properties are available in Appendix A.

4 Experiments

Multimodal industry classification task tests MLLM using various resources to predict economic activities in two pipelines: Zero-Shot and

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

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

Multi-Turn. Zero-Shot detects NACE section in single inference using multimodal inputs. Multi-turn has clue extracting agents for each input type, and a decision-making agent processes these clues.

We have several experiment dimensions for model selection, prompting strategies, input configurations, and output structures. Using frequency vectors in the clue analysis stage, we quantify intermediate agent effectiveness and correctness. We propose new metric to analyze model uncertainty.

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: Multi-turn has two stages: Clue Extraction and Decision Making. The clue extraction contains agents designed to generate *clues* up to the number of inputs (e.g., OSM). Decision-making agent uses intermediate agent responses, *clues*, and the entity name to choose NACE sections.

4.2 Experiment Dimensions

Models: We selected open and closed-source models: 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), Gemini 2.5 (Gemini 2.5 Team, 2025), GPT 5 - Mini, and GPT 5.1 (OpenAI, 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 include NACE section codes and titles. `Extended` prompt has section summaries from official guidelines with description, their content, and exclusions. Templates and contents are given in the Appendix E.

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.

4.3 Clue Analysis

In our multi-turn pipeline, clue agents are instructed to process specific input types to generate free-form texts with keywords, Table 8, describing economic activities. For example: [retail] for section G (Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles). 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).

Upon grouping keywords by sections, we obtain normalized keyword counts. We refer to this scaled column vector as the frequency vector, $v_{i,c}$. In the 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:

$$\begin{matrix}\text{Ground Truth: } v_{g_i}[i] \\ \text{Prediction: } v_{p_i}[i]\end{matrix} = \begin{bmatrix} v_{i,c=\text{OSM}}[g_i \mid p_i] \\ v_{i,c=\text{Satellite}}[g_i \mid p_i] \\ v_{i,c=\text{Wikidata}}[g_i \mid p_i] \\ v_{i,c=\text{Wikipedia}}[g_i \mid p_i] \\ v_{i,c=\text{Website}}[g_i \mid p_i] \end{bmatrix} \tag{1}$$

In the example Figure 3, we use index K for the ground truth vector. For OSM, $v_{\text{OSM}}[K]$, satellite $v_{\text{Satellite}}[K]$, 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, $v_p = [12/13; 1/3; 0; 0; 0]$.

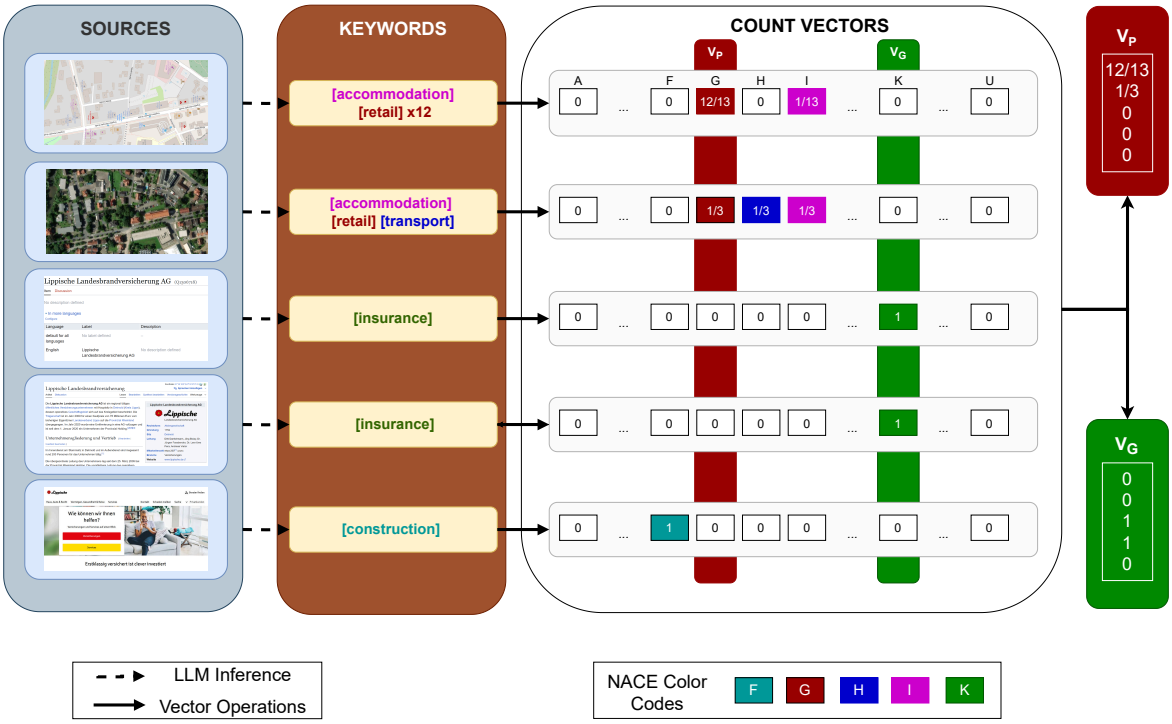


Figure 3: Example with ground truth NACE sections K and prediction G. Clues and predictions are obtained via InternVL-3 8B. Clue Analysis Methodology: (1) Sources are forwarded into separate MLLM Agents for clue extraction. (2) Keywords based on predefined Economic Activities are extracted and grouped 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).

4.4 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:

$$\text{Unknown Ratio (UR)} = U/I \tag{2}$$

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

- **Correctness:** Measures relatedness of clues 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.

$$\text{Correctness}_c = \frac{\sum_i^{I_c} v_{g_i}[i, c]}{I_c} \tag{3}$$

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

- **Effectiveness:** Measures how clues 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.

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

  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 demonstrated the baseline results in Table 2 using the Zero-Shot pipeline, `Simple` prompt, and `Text` output. The name of the entity is given for every input configuration.

InternVL 3-78B and GPT 5-Mini achieved the highest performance for open and closed-source MLLMs. InternVL3-14B is the best-performing small ($\leq 14B$) model. Due to its limited context window and weaker performance, we exclude

Model	Size (B)	None	Satellite	External	Satellite + OSM	Satellite + External	All
Open Source Models							
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**
Closed Source Models							
Gemini 2.5 Flash		58.40	63.50	71.00	66.80	73.80	72.40
GPT 5 Mini		**62.00**	**66.80**	**71.90**	**68.90**	**74.10**	**73.30**
GPT 5.1		57.80	59.60	69.10	63.40	70.00	70.20

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 for the open-source model, and the best performance among closed-source models. GPT 5 Mini and InternVL 3-78B are the best-performing closed and open source models.

LLaVA v1.6 from the remaining experiments. The difficulty of the task is apparent as even 70B+ models failed to reach **65%** accuracy. Also, the best open source performance is on par with the best model's name-only performance, which indicates the gap between open and closed source MLLMs.

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

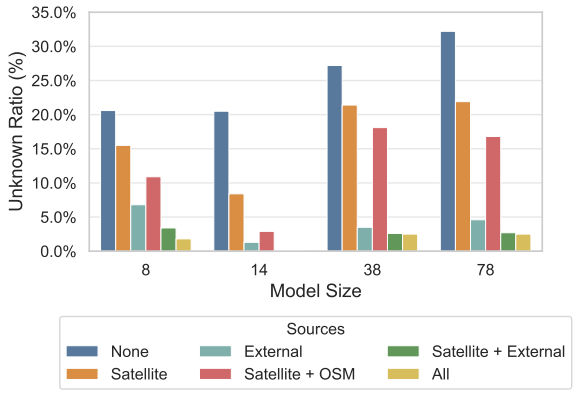


Figure 4: Baseline unknown ratio of InternVL 3 for input configurations. Unknown ratio corresponds to UNK responses over all inferences.

To quantify uncertainty of the InternVL 3 responses, we listed unknown ratios in Figure 4. 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. Providing additional inputs yields accuracy gains of at most **20%**, making the entity name the strongest predictive signal. Furthermore, model performance is the best when external resources are given in context compared to geospatial resources.

Unlike accuracy, the unknown ratio reveals that the name alone is not enough for a robust prediction. We also noted that the 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. For open-source models, we selected one small (≤ 8) and one large ($\geq 30B$) model for InternVL 2.5 and 3 and QwenVL 2.5.

Model	Size	Baseline	Explanation	Extended Prompt	Multi-Turn	Extended Prompt +	Mixture
Open Source Models							
InternVL 2.5	4B	6.40	23.10 (16.70)	8.30 (1.90)	22.00 (15.60)	10.60 (4.20)	**29.20 (22.80)**
	38B	60.10	61.80 (1.70)	64.20 (4.10)	58.20 (-1.90)	**65.00 (4.90)**	60.20 (0.10)
InternVL 3	8B	41.70	42.30 (0.60)	36.90 (-4.80)	**49.80 (8.10)**	38.00 (-3.70)	45.70 (4.00)
	38B	58.30	59.80 (1.50)	61.30 (3.00)	61.60 (3.30)	**64.10 (5.80)**	62.60 (4.30)
QwenVL 2.5	7B	23.10	30.50 (7.40)	27.30 (4.20)	38.90 (15.80)	31.00 (7.90)	**45.90 (22.80)**
	32B	56.40	60.00 (3.60)	60.40 (4.00)	55.70 (-0.70)	**65.40 (9.00)**	62.00 (5.60)
Closed Source Models							
Gemini 2.5 Flash		72.40	72.50 (0.10)	**74.30 (1.90)**	71.20 (-1.20)	74.00 (1.60)	72.70 (0.30)
GPT 5 Mini		73.30	74.00 (0.70)	**74.70 (1.40)**	74.30 (1.00)	72.90 (-0.40)	74.20 (0.90)
GPT 5.1		70.20	69.40 (-0.80)	69.70 (-0.50)	69.00 (-1.20)	68.50 (-1.70)	**70.40 (0.20)**

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**.

The accuracy results are shown in Table 3. For these experiments, we used all the inputs available.

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 ($\geq 30B$) and proprietary models, we obtained the best performances without multi-turn. The best performing smaller model is also the most recent model, InternVL 3-8B.

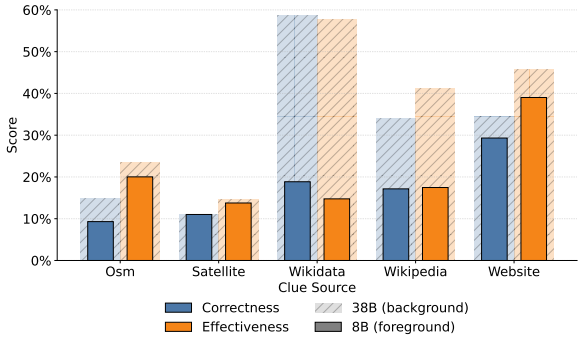


Figure 5: InternVL 3 (8B and 38B) correctness and effectiveness scores for each input. In these experiments, multi-turn pipeline with extended prompt and classification explanations is used.

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 with respect to ground truth. We demonstrated InternVL 3 (8B and 38B) results for the mixture configuration (multi-turn with extended prompt and classification explanation) in Figure 5. Other model results are available in Appendix Table 11.

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. Except for websites, the 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 with 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 is mostly present in or similar to the pretraining data, (2) the used models are not adapted to remote sensing.

5.4 Ablations

Qualitative Results: In Table 4, we selected examples from MONETA containing satellite images and websites (translated and summarized). In these examples, the generated clues contradict each other. While websites are often the most effective source, they may emphasize sales-related information, introducing a bias toward NACE Section G (Wholesale and Retail Trade). When websites are absent or less informative, satellite imagery can instead enable correct identification of the industry.

However, as the quantitative results show for the

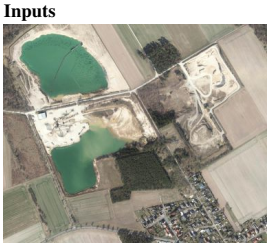
Example 1	Example 2
Inputs	**Inputs**
	
Kieswerk Bahrdorf is a producer and wholesaler of bulk materials such as sand and gravel, supplying the greater Wolfsburg area.	AnconAmbiente provides urban waste collection and environmental services for municipalities in the Province of Ancona.
Clues	**Clues**
Satellite: [quarrying] Excavation and heavy machinery.	**Satellite:** [manufacturing] Large industrial buildings.
Website: [wholesale] Producer and wholesaler of bulk materials.	**Website:** [waste] Waste collection services.
Rationale and Label	**Rationale and Label**
Satellite imagery shows excavation and heavy machinery consistent with quarrying. The term "Kieswerk" explicitly refers to a gravel pit.	Observed infrastructure and website content indicate organized waste collection rather than manufacturing.
Label: Mining and Quarrying	**Label:** Water Supply; Sewerage, Waste Management and Remediation Activities

Table 4: Qualitative MONETA examples for correct inferences with contradicting clues Satellite image and website summary are followed by extracted clues. LLM clues, rationale, and decision are obtained via InternVL 3-38B in the mixture configuration. Green indicates content supporting ground truth while orange contents do not match the ground truth label.

second example, visual cues can also be misleading. For a robust and accurate industry classification, both text clues and visual clues should be used.

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 5: 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 shown in Appendix C. As they fine-tuned models, we included test results for few shots (1 sample per class) and adapted models (with LORA) in Table 5. We used Qwen 2.5 text model as it is text core for both QwenVL 2.5 and InternVL 3.

With zero-shot pipeline, we observed similar performances for both datasets. We had minimal gains with few-shot for the 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 to a fixed classification scheme makes models fragile to future revisions. To demonstrate this, we evaluated adapted models on an alternative task and observed a performance drop exceeding **35%** relative to zero-shot inference. Not only does fine-tuning require a significant amount of labeled data, but fine-tuned models are also not usable after classification schemes change. As noted by Guo et al. (2025), industry classifications have changed in form multiple times over the years. In contrast, our prompting strategy remains adaptable and robust to such updates.

6 Conclusion

In this work, we introduce MONETA, a new dataset and task for multimodal industry classification. Our benchmark reflects real-world challenges in business registers by enabling industry classification using satellite and OSM imagery with external resources and NACE labels.

We proposed a multi-turn pipeline that generates clues from each resource and introduces metrics for their quantitative analysis. We validated our pipeline on two existing unimodal datasets and outperformed one configuration. Our experiments highlighted the limitations of fine-tuned models in cross-domain settings and the robustness of our zero-shot alternative. MONETA reveals the difficulty of the task and the textual bias of MLLMs.

Beyond our methodology, MONETA is relevant to policymakers and financial experts by supporting financial risk assessment, market analysis, and regional economic monitoring. We enable fast and reliable industry identification for newly founded or data-sparse entries in business registers. Future work will analyze its integration into real-world decision-making.

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, and 8 others. 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, and 1 others. 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.

Gemini 2.5 Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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 preprint*. 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.

Qiang Li, Mingkun Tan, Xun Zhao, Dan Zhang, Daoan Zhang, Shengzhao Lei, Anderson S. Chu, Lujun Li, and Porawit Kamnoedboon. 2025. How LLMs react to industrial spatio-temporal data? assessing hallucination with a novel traffic incident benchmark dataset. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 36–53, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

OpenAI. 2025. Gpt-5. `https://openai.com`. Large language model.

Maryan Rizinski, Andrej Jankov, Vignesh Sankaradas, Eugene Pinsky, Igor Miskovski, and Dimitar Trajanov. 2023. Company classification using zero-shot

learning. *arXiv preprint*. 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 preprint*. 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, and 3 others. 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 preprint*. 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 preprint*. 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

MONETA contains 1,000 businesses in Europe with EU Guidelines' NACE economic activity labels. Each entry contains two geospatial and at least one textual resource. As the NACE section T, Activities of Households as Employers; Undifferentiated Goods- and Services-producing Activities of Households for Own Use, cannot be obtained through OSM, we used the remaining 20 NACE sections for economic activities. For each section, we list 50 entities.

In Table 6, 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.

A sample entry of MONETA used in qualitative analysis has the attributes given in Table 7. This entry contains only the website as the external source. Due to the content length, we provided the link instead. We replaced image paths with placeholders.

A.1 NACE to OSM Mapping

OSM contains tags describing the geospatial entity. These tags can indicate contact information,

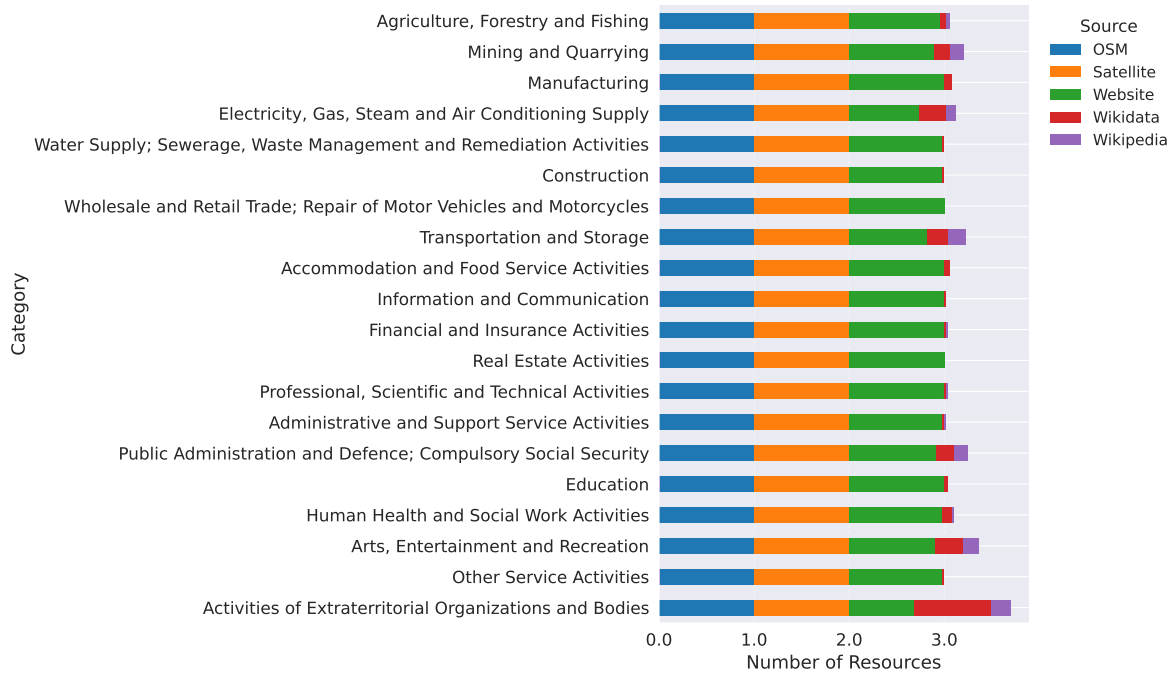


Figure 6: Average number of associated resources (OSM images, satellite images, website text, Wikidata, and Wikipedia) per NACE sections in the MONETA dataset

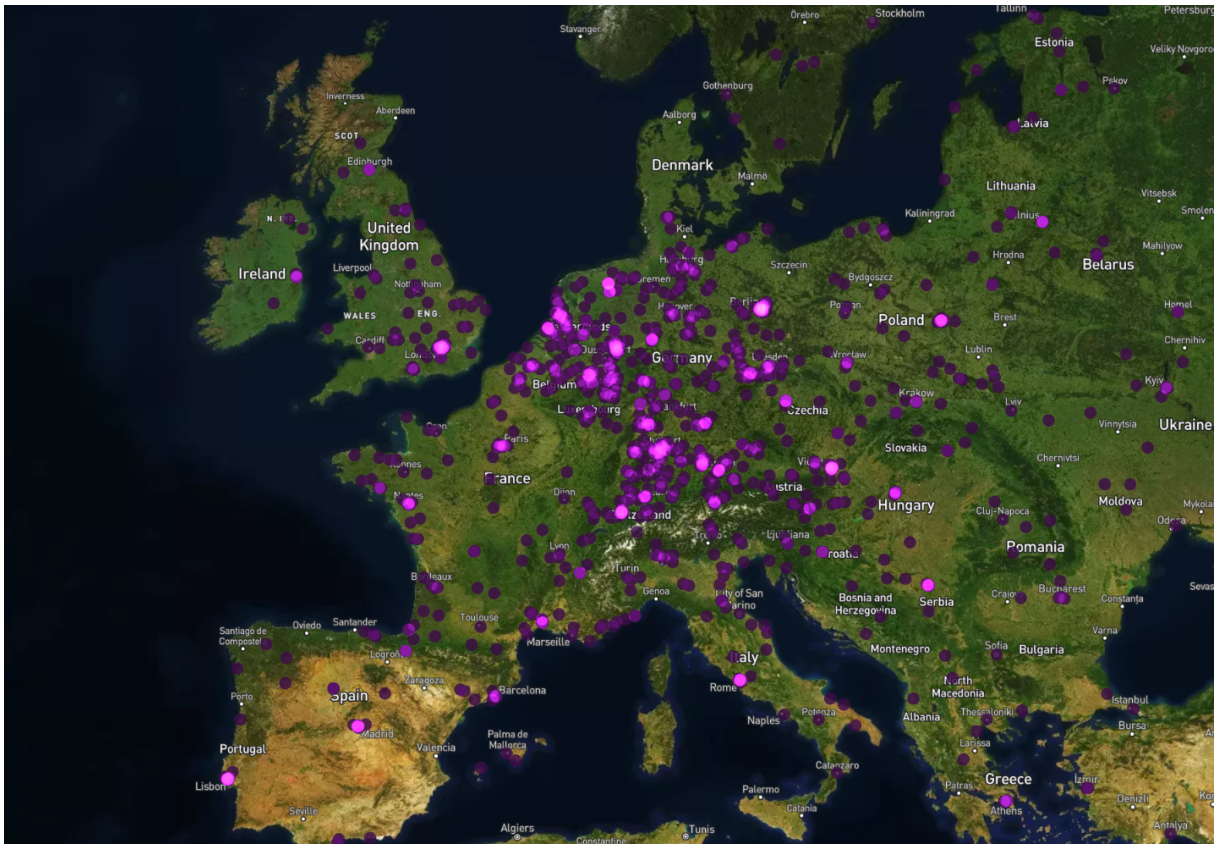


Figure 7: Spatial distribution of MONETA entries across Europe, showing the geographic coverage of OSM-derived entities used for NACE classification.

address, external database references (such as Wikidata), building structures, and many economic

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 6: 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.

Attribute	Value
id	122563530
type	way
name	Heim Kieswerk
bbox	[12.4893727, 50.9761359, 12.5089029, 50.9916218]
category	B (Mining and quarrying)
osm_tags	addr:city = Nobitz addr:country = DE addr:housenumber = 14c addr:postcode = 04603 addr:street = Altenburger Straße landuse = quarry resource = sand operator = Heim Kieswerk Nobitz GmbH & Co. KG
image_paths	OSM: OSM_PATH.png Satellite: Satellite_PATH.png
sources	Website text extracted from https://www.heim-gruppe.de Wikipedia: – Wikidata: –

Table 7: MONETA entry for *Heim Kieswerk* derived from OpenStreetMap and external sources.

properties. However, there is no one-to-one mapping connecting OSM to any existing industry classification framework. As far as we are concerned, this study is the first connecting OSM tags to the NACE industry classification scheme.

In this section, we illustrate the data preparation workflow shown in Figure 2 using NACE Rev.2 Section K as an example. We first extract the official NACE guideline from the RDF/XML source. Extracted fields are title, content, scope, additional content, and exclusions as shown in Figure 8.

These textual descriptions are then provided to Gemini to generate a candidate list of OpenStreetMap (OSM) tags relevant to the economic activities defined under Section K (Figure 9). As Gemini can hallucinate the tags or recommend rarer tags in the OSM database, we do not use the generations directly. Instead, we verify the generated tags list, one by one, to ensure they exist and fit the scope of the related economic activity. We discard the tags that are non-standard, weakly related, or rarely used. We also add relevant tags from the OSM TagInfo database matching the NACE description.

For example, company=insurance is a rare tag with 27 entries around the globe. office=company, on the other hand, is a broad tag that can correspond to any company that may not have an economic activity related to Financial and Insurance activity. office=financial is a valid and common tag fitting to section K. Furthermore, it is often used with address tags, which helps the data quality assurance.

After creating the NACE to OSM Mapping, we retrieve the elements from OSM's most recent European extract. To ensure data quality, we limit the search to entries with a name tag in OSM. We iteratively add filters to generate versions of our datasets. Our bronze variant has the simplest filter with NACE-induced tags and the name filter. Upon this version, we include the terms with valid address tags and prepare our silver dataset. Then, we finally include external database pointer tags for Wikidata, Wikipedia, and the Website to generate the gold version. MONETA contains 50 entries per category from this gold version.

A.2 NACE Details

We extracted NACE codes, titles, descriptions and keywords for prompting. Summary of these attributes are given in Table 8.

B MONETA-10K

Our NACE to OSM mapping allows us to retrieve elements for more than the 1,000 businesses we used in this study. However, due to computational resources and budgeting, we can test various input configurations and experiment dimensions with the released version of MONETA. Therefore, we sampled 50 entries per NACE section. However, using the same mapping, we also generated a more comprehensive benchmark which we call MONETA-10K. This benchmark, as the name implies, contains 10,000 businesses with NACE section labels.

NACE Rev.2 RDF/XML Extract — Section K

Source

Retrieved from:
https://publications.europa.eu/resource/authority/ux2/nace2/K

Parsed Content

```
{
  '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 holding companies and trusts,
  funds and similar financial entities.',
  'Exclusion': None
}
```

Figure 8: RDF/XML NACE official guideline content for Section K (Financial and Insurance Activities).

Gemini-Generated OSM Tags Section 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 OpenStreetMap tag candidates for the NACE Financial and Insurance Activities section (K). Relevant tags align with official NACE definitions, while other tags are either rare, non-standard, or weakly related.

B.1 Dataset Details

In this version, all entries possess at least 1 external resource and 2 geospatial images. The distribution of sections is given in Figure 10. Furthermore, we provided details of external resources in Table 9.

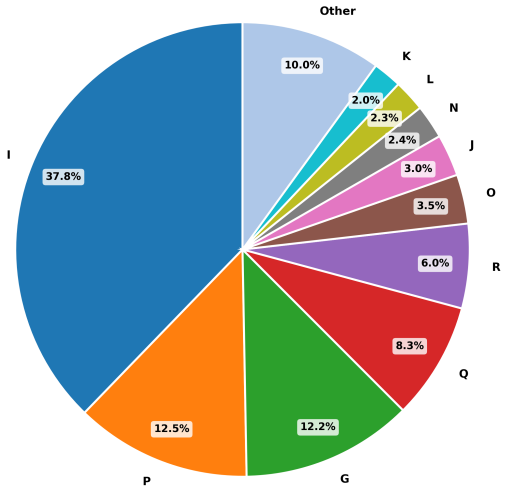


Figure 10: NACE section distribution for MONETA-10K. Sections less than 2% are grouped into other for demonstration.

B.2 Comparison with MONETA Results

We examined MLLM performance on MONETA-10K using the InternVL 3 - 8B model in our baseline configuration. We used text output with single token inference and provided only the NACE sections and titles in the system prompt. We used all available resources per entry.

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 8: NACE Section Codes, Titles, AI-generated descriptions and keywords (from official guidelines). During the multi-turn inference, MLLMs will generate economic activity clues based on provided keywords.

In Table 10, we demonstrated results for macro and weighted F1-Score and Precision, Recall. The results for MONETA and MONETA-10K differ less than 5% which indicates that MONETA can be used in NACE-based industry classification like a larger counterpart.

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

Text Source	Entry Count
Website	9,015
Wikidata	276
Wikipedia	13
Wikidata + Website	315
Wikidata + Wikipedia	147
Wikipedia + Website	13
All	221

Table 9: MONETA-10K external resource counts. All denotes the existence of Wikidata, Wikipedia, and Website

Metric	MONETA-10K	MONETA
Macro F1-score	39.40	38.70
Weighted F1-score	45.90	40.70
Precision	52.30	52.30
Recall	39.40	39.70

Table 10: Macro and Weighted F1, recall, and F1-score for the MONETA-10K and MONETA datasets with InternVL 38B using Zero-Shot pipeline, Text output, Simple prompt and all available resources.

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 Section-wise Analysis

Based on the available configurations (Baseline, Explanation, Extended Prompt, Multi-Turn, Extended Prompt + (Explanation), Mixture), we created confusion matrices between ground truth and prediction results. We retrieved the counts from the diagonals and visualized them with respect to configurations in Figure 11. The smaller models performed poorly for the sections M (Professional, Scientific, and Technical Activities) and S (Other Service Activities). The usage of multi-turn allowed smaller models to detect B (Mining and Quarrying) and U (Activities of Extraterritorial Organisations and Bodies). Larger models are overall consistent with the exceptions of sections F (Construction), N (Administrative and Support Service Activities), and S (Other Service Activities). The effect of prompt context enrichment and multi-turn is also visible for U (Activities of Extraterritorial Organisations and Bodies) for larger models.

D.2 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.

D.3 Clue Ablations

Model section preferences Using the keywords in the freeform text, we grouped clue contents into NACE sections using the keyword list. Resulting groupings formed the clue frequency vectors. From clue frequency vectors, we identified percentages for each NACE 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.

We also analyzed the obtained clues using our correctness and effectiveness measures in Table 11. For these experiments, we used open-source MLLMs InternVL 2.5 (4B and 38B), InternVL 3 (8B, 14B, 38B, and 78B) and QwenVL 2.5 (7B, 32B, and 72B). For the larger models, we observed that the highest correctness and effectiveness are attained via Wikidata context. The smaller models can utilize website context the most and thus achieve their highest effectiveness and correctness

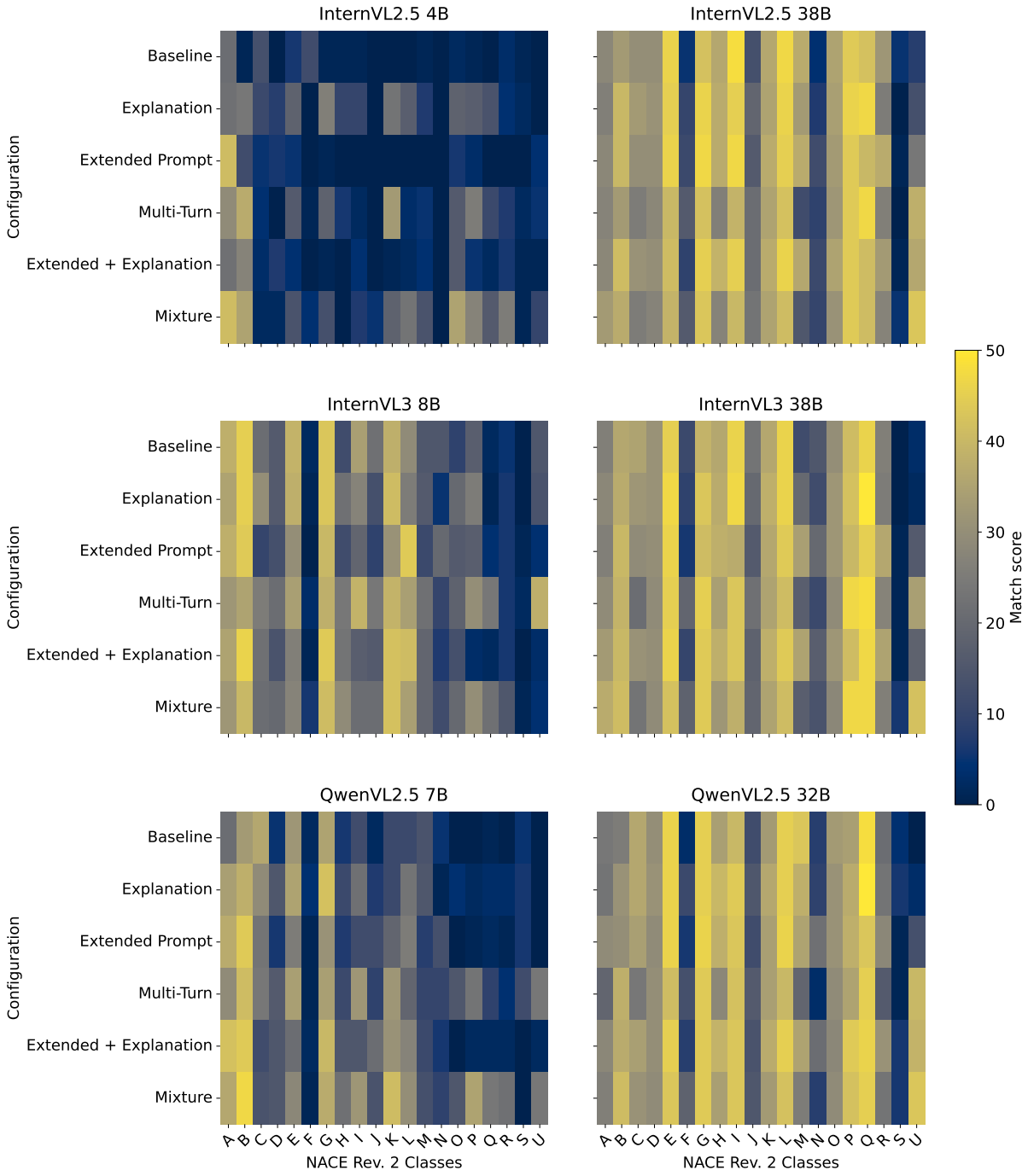


Figure 11: NACE Section-wise analysis for InternVL 2.5 (4B and 38B), InternVL 3 (8B and 38B), and QwenVL 2.5 (7B and 32B). Rows indicate experiment configurations. Mixture denotes Multi-Turn pipeline with classification explanations and prompt enrichment. Columns are the NACE section letters given in Table 8.

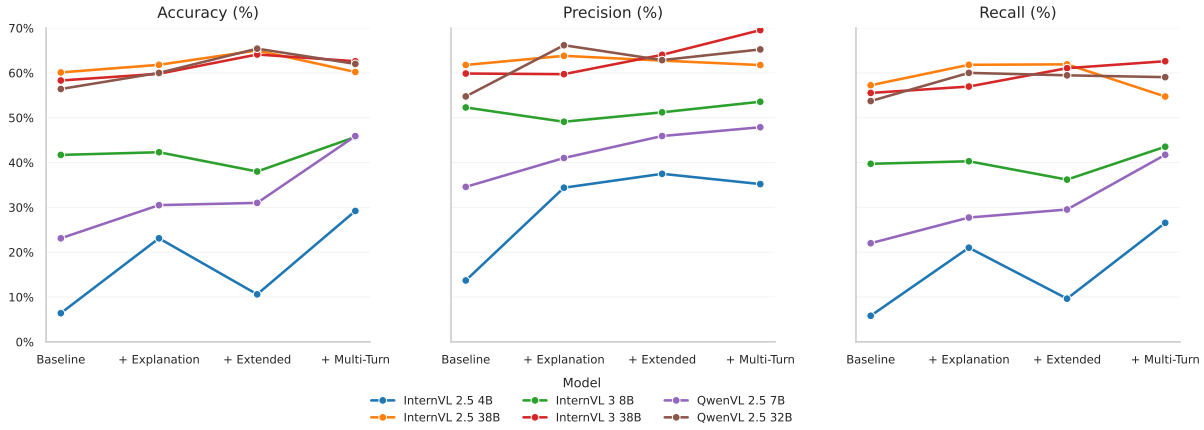


Figure 12: Experiment configuration results given in Accuracy, Precision, and Recall for InternVL 2.5 (4B and 38B), InternVL 3 (8B and 38B), and QwenVL 2.5 (7B and 32B). At each setting, the given configuration is added. The order of configurations is: Baseline, Explanation, Explanation + Extended Prompt, Mixture (with Multi-Turn).

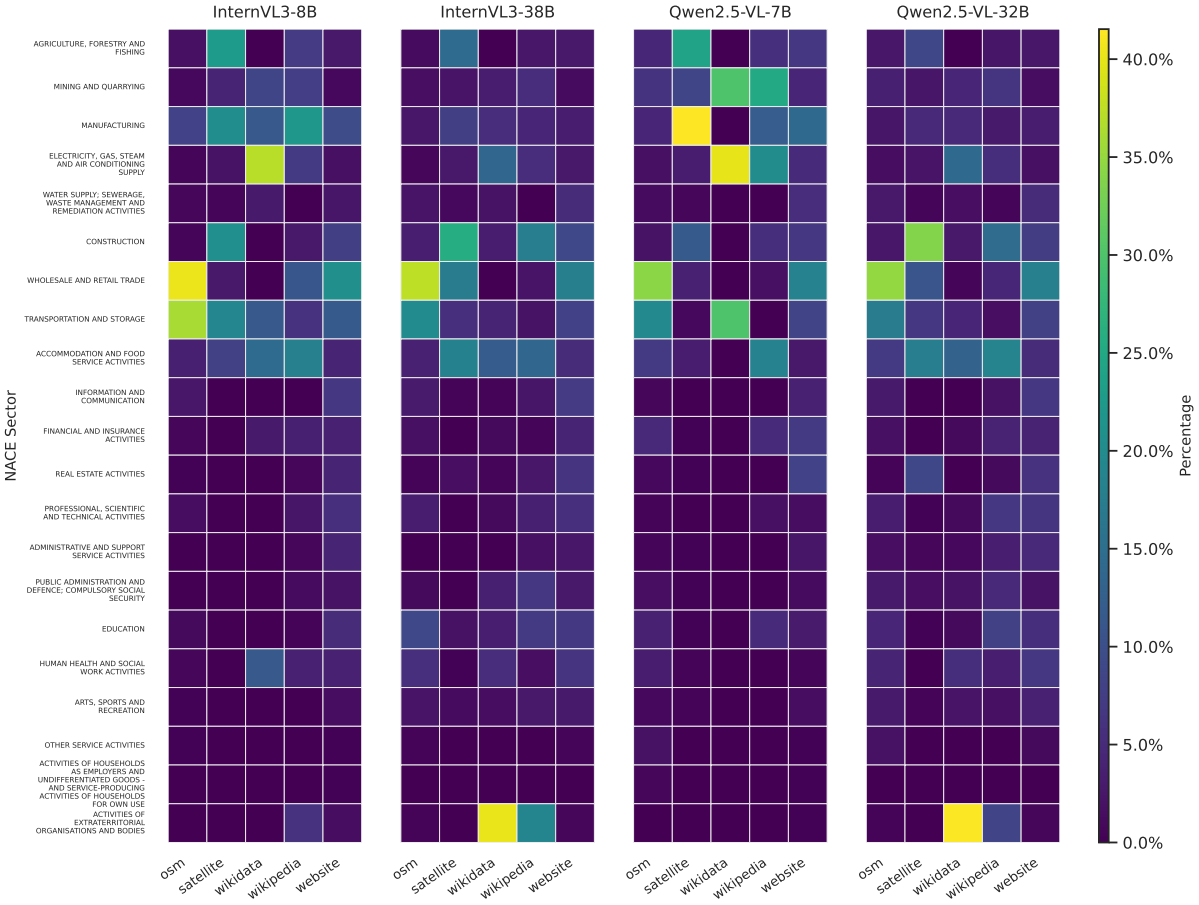


Figure 13: Clue Keywords confusion matrices obtained for InternVL 3 (8B and 38B), and QwenVL 2.5 (7B and 32B). Extracted keywords are grouped based on NACE sections in the rows. The columns are resources: OSM, Satellite, Wikidata, Wikipedia and Website.

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**
	14	10.22	19.55	7.83	9.02	**40.57**	**47.95**	31.25	37.11	34.23	45.69
	38	14.77	23.40	11.04	14.62	**58.61**	**57.65**	33.92	41.23	34.58	45.72
	78	10.58	13.56	9.73	11.34	**57.38**	**55.33**	35.08	34.62	39.18	51.04
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
	72	16.28	22.32	7.22	9.95	**50.82**	**50.00**	22.57	23.21	27.40	35.10

Table 11: Performance comparison in the Mixture setting, reporting correctness (left) and effectiveness (right) for each model across input modalities. Image-based inputs are highlighted. **Bold** indicates the best performance in correctness/effectiveness for the model and size pair.

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**
	14	67.90	44.10	73.77	82.69	**90.65**
	38	87.10	78.40	78.69	**90.38**	79.38
	78	73.40	30.00	74.59	80.77	**85.76**
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
	72	**73.20**	30.50	60.66	53.85	68.33

Table 12: Information Discovery of inputs. Image inputs are highlighted. **Bold** denotes the highest information discovery from a model-size pair.

scores when using these data. Among the visual clues, OSM images appear to be more useful compared to satellite images. The clue effectiveness and correctness indicate that even for the best resource, Wikidata, metric performances are below 60%. Thus, industry classification cannot be solved using only a single source.

D.4 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:

$$\text{Information Discovery } (ID_c) = 1 - \frac{NEI_c}{I_c} \quad (5)$$

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 12, Information Discovery Rates are given. Smaller models 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. Among the text sources, we observed that Website and Wikidata to contain more information regarding economic activities. Especially, InternVL 3 models generated economic activity clues for more than 80% of the examples. According to the results, InternVL 3 has a stronger vision core compared to its predecessor, as it is now able to utilize satellite and OSM images more than **30%** at the same size.

E Prompts

E.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.

NACE–OSM Tag Mapping Prompt

Task Description
You will be given a description of a **NACE code**, representing a business activity. Your task is to identify relevant **OpenStreetMap (OSM)** tags that can be used to classify businesses or locations corresponding to this activity.

NACE Code Description:

{RDF/XML Extract}

Response Format
Your response must consist **only** of a Python list of OSM tags, where each element is a string in the form `key=value`.

```
["landuse=retail",
"shop=supermarket",
"amenity=parking"]
```

Constraints

- Every tag must include an = sign (e.g., `shop=supermarket`)
- Do **not** include bare keys such as `shop` or `amenity`
- Do **not** include explanations or additional text
- Do **not** include Python code markers
- Do **not** use tags unrelated to business activities (e.g., `landuse=forest`)
- Output **only** the Python list

OSM Tags:

E.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.

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.
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

E.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 the model to use the name if no input is provided. Then, we provide the output prompt depending on the experiment setup. Finally, we give the context based on the input configuration.

E.4 Multi Turn

In our multi-turn pipeline, we have intermediate-level agents for each input type (Satellite, OSM,

Zero-Shot NACE Classification Prompt

Role
You are an assistant designed to identify *economic activities* from heterogeneous geospatial and textual resources.
Inputs

- **Images**: OpenStreetMap (OSM), Satellite imagery
- **Textual**: Wikidata, Wikipedia, Website
- **Entity name**

Visual Analysis (Images) Identify relevant geospatial features, including but not limited to:

- Buildings
- Terrain
- Streets

Contextual Analysis (Text) Extract economic context such as:

- Products and services
- Activities
- Business type
- Industry

Task
Based on the extracted attributes and the entity name, predict the corresponding **NACE Rev.2 economic activity sector code**.

`{NACE_CONTEXT}`

Available Resources

- `osm`: OSM image
- `satellite`: Satellite image
- `source`: Wikidata / Wikipedia / Website

If no external resources are provided, rely solely on the entity name.
Output Format

`{OUTPUT_FORMAT}`

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 8. These keywords are expected in the output for easier grouping of the free-form text. We provided a sample prompt containing shared instructions here.

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

Clue Extraction Agent
Shared Instructions

You are an agent tasked with extracting *explicit economic activity clues* from a single information source.

General Rules

- Only extract activities with **direct textual or visual evidence**.
- The provided keyword list defines all valid economic activity categories.
- Match only exact keywords or clear synonyms.
- Do **not** infer, guess, or generalize beyond the source.
- When mentioning an activity, wrap it in `[ ]` exactly as in the keyword list.
- For every activity, cite the exact supporting feature, tag, phrase, or entity.
- If no activity is present, output exactly: `"No economic activity clues found."`
- Output language must be English.
- Maximum output length: 512 tokens.

Output Format

Economic activity clues:
- [keyword] supporting evidence from the source

Multi-Turn Classification Prompt

Role

You are an assistant designed to identify *economic activities* from multiple, incremental information sources.

Inputs

You may be provided with clues from the following sources:
Wikidata, Wikipedia, Websites, OpenStreetMap (OSM) images, Satellite images

Task

Based on the provided clues and the entity name, identify the corresponding **NACE economic activity sector code**.

`{NACE_CONTEXT}`

Note that you may not be given all of the clues. If no clues are provided, rely solely on the entity name.

Output Format

`{OUTPUT_FORMAT}`

E.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 randomly selected with a fixed seed one example per class from the training set. The prompt structures are identical for each task. It starts with a set of instructions followed by available categories and choices.

Ablation Instructions

Classify the company into one industry sector.
You are given codes and titles.
Respond with EXACTLY ONE UPPERCASE LETTER.
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: [Available options]

ExioNAICS Prompt

`{ABLATION_INSTRUCTIONS}`

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

Company Websites Prompt

`{ABLATION_INSTRUCTIONS}`

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