MVT Appendix Materials

S1: Land-Use Taxonomy and Hierarchical Mapping

The following specifies the hierarchical mapping between first-level land-use categories and their corresponding second-level subcategories, forming a structured land-cover taxonomy. The taxonomy is constructed according to the national land-use classification standard of the People’ s Republic of China, as described in Bai-ming Chen *et al.* (2007), and provides a consistent semantic hierarchy for land-use interpretation.

Label Taxonomy

```
LAND_USE_TAXONOMY: Dict[str, Dict] = {
  "01": {
    "level1_name": "Cultivated Land",
    "level2": {
      "011": {"level2_name": "Paddy Field"},
      "012": {"level2_name": "Irrigated Land"},
      "013": {"level2_name": "Dry Land"}
    }
  },
  "02": {
    "level1_name": "Garden Land",
    "level2": {
      "021": {"level2_name": "Orchard"},
      "022": {"level2_name": "Tea Garden"},
      "023": {"level2_name": "Other Gardens"}
    }
  },
  "03": {
    "level1_name": "Forest Land",
    "level2": {
      "031": {"level2_name": "Forest"},
      "032": {"level2_name": "Shrubland"},
      "033": {"level2_name": "Other Forest Land"}
    }
  },
  "04": {
    "level1_name": "Grassland",
    "level2": {
      "041": {"level2_name": "Natural Grassland"},
      "042": {"level2_name": "Artificial Grassland"},
      "043": {"level2_name": "Other Grassland"}
```

```
      }
    },
    "05": {
      "level1_name": "Commercial Service Land",
      "level2": {
        "051": {"level2_name": "Retail Land"},
        "052": {"level2_name": "Accommodation and Catering Land"},
        "053": {"level2_name": "Business and Financial Land"},
        "054": {"level2_name": "Other Commercial Land"}
      }
    },
    "06": {
      "level1_name": "Industrial, Mining, and Storage Land",
      "level2": {
        "061": {"level2_name": "Industrial Land"},
        "062": {"level2_name": "Mining Land"},
        "063": {"level2_name": "Storage Land"}
      }
    },
    "07": {
      "level1_name": "Residential Land",
      "level2": {
        "071": {"level2_name": "Urban Residential Land"},
        "072": {"level2_name": "Rural Homestead Land"}
      }
    },
    "08": {
      "level1_name": "Public Administration and Public Service Land",
      "level2": {
        "081": {"level2_name": "Government Land"},
        "082": {"level2_name": "Press and Publication Land"},
        "083": {"level2_name": "Scientific and Educational Land"},
        "084": {"level2_name": "Medical and Charity Land"},
        "085": {"level2_name": "Cultural Service Land"},
        "086": {"level2_name": "Public Facilities Land"},
        "087": {"level2_name": "Park and Green Space"},
        "088": {"level2_name": "Scenic and Natural Heritage Land"}
      }
    },
    "09": {
      "level1_name": "Special Land",
      "level2": {
        "091": {"level2_name": "Military Facilities Land"},
        "092": {"level2_name": "Embassies and Consulates Land"},
        "093": {"level2_name": "Prison Land"},
        "094": {"level2_name": "Religious Land"},
        "095": {"level2_name": "Cemetery Land"}
```

```
      }
    },
    "10": {
      "level1_name": "Transportation Land",
      "level2": {
        "101": {"level2_name": "Railway Land"},
        "102": {"level2_name": "Road Land"},
        "103": {"level2_name": "Street Land"},
        "104": {"level2_name": "Rural Road Land"},
        "105": {"level2_name": "Airport Land"},
        "106": {"level2_name": "Port Land"},
        "107": {"level2_name": "Pipeline Transportation Land"}
      }
    },
    "11": {
      "level1_name": "Water Bodies and Hydraulic Facility Land",
      "level2": {
        "111": {"level2_name": "River Surface"},
        "112": {"level2_name": "Lake Surface"},
        "113": {"level2_name": "Reservoir Surface"},
        "114": {"level2_name": "Pond Surface"},
        "115": {"level2_name": "Coastal Tidal Flats"},
        "116": {"level2_name": "Inland Tidal Flats"},
        "117": {"level2_name": "Ditches"},
        "118": {"level2_name": "Hydraulic Construction Land"},
        "119": {"level2_name": "Glacier and Permanent Snow"}
      }
    },
    "12": {
      "level1_name": "Other Land",
      "level2": {
        "121": {"level2_name": "Idle Land"},
        "122": {"level2_name": "Facility Agricultural Land"},
        "123": {"level2_name": "Ridge Land"},
        "124": {"level2_name": "Saline-Alkali Land"},
        "125": {"level2_name": "Swamp"},
        "126": {"level2_name": "Sandy Land"},
        "127": {"level2_name": "Bare Land"}
      }
    }
  }

```


For example, **Cultivated Land (Level-1 ID: 01)** is a first-level category, which contains three Level-2 subcategories: **Paddy Field (011)**, **Irrigated Land (012)**, and **Dry Land (013)**. Each Level-

2 category inherits the semantic scope of its corresponding Level-1 category, forming a deterministic and interpretable coarse-to-fine land-use hierarchy.

S2: MLLM Finetuning Prompt

S2.1 First-Step Preprocessing

First-Step Category Mapping

```
# Fixed category order -> numeric labels (1..29)
CATEGORY_ORDER = [
    "BareLand","BaseballField","Beach","Bridge","Center","Church","Commercial",
"DenseResidential","Desert","Farmland","Forest","Industrial","Meadow","MediumResidential",
"Mountain","Park","Parking","Playground","Pond","Port","RailwayStation","Resort","River",
    "School","SparseResidential","Square","Stadium","StorageTanks","Viaduct"
]
```

Before first-step fine-tuning, we **anonymize all semantic class names** by mapping them to a **fixed, deterministic integer label set** [1,29]. Concretely, we define a constant `CATEGORY_ORDER` and assign labels sequentially according to this order (Specifically, BareLand is labeled as 1, BaseballField as 2, and so on, up to Viaduct as 29 following the given order.). **This anonymization is completed prior to any prompt construction or model training**.

S2.2 First-Step Finetuning

We apply the following prompt structure as our first-step finetuning period.

First-Step Finetuning Prompt

```
PROMPT = (
    "<image>\n"
    "You are a senior remote-sensing analyst. Carefully examine the remote-sensing RGB image at multiple scales and infer its numeric label ID (1-29) strictly from visual evidence.\n"
    "Internally consider: global spatial layout and landform; geometry, size, and alignment of man-made structures; texture repetitiveness and granularity; surface/material cues; linear networks (roads, tracks, embankments, shorelines); density and relative scale indicators; color/tonal/spectral
```

```
    contrasts; cast shadows and illumination; contextual boundaries and
    transitions.\n"
5       "Open-set constraint: NEVER state or imply any category or scene name, and
    do not verbalize your rationale.\n"
6       "Output format: return ONLY the integer label in [1, 29] as plain text
    with no extra words, symbols, or punctuation."
7   )
8
9   INPUT_TXT = (
10      "Return one integer in [1, 29]. Do not include any words, labels, or
    explanations."
11  )
```

S2.3 Second-Step Preprocessing

Second-Step Category Mapping

```
1   Category Mapping List
2   <original_class_name> → <anonymized_category_id> (number_of_images)
3
4   airfield → category0001 (800)
5   airplane → category0002 (800)
6   airport → category0003 (700)
7   avenue → category0004 (1031)
8   bare_land → category0005 (1137)
9   baseballdiamond → category0006 (800)
10  basketball_court → category0007 (700)
11  beach → category0008 (2118)
12  bridge → category0009 (1136)
13  buildings → category0010 (1117)
14  chaparral → category0011 (800)
15  church → category0012 (700)
16  circular_farmland → category0013 (700)
17  city_road → category0014 (367)
18  coastline → category0015 (653)
19  commercial_area → category0016 (700)
20  container → category0017 (1166)
21  crossroads → category0018 (803)
22  dam → category0019 (434)
23  denseresidential → category0020 (1600)
24  desert → category0021 (2000)
25  farmland → category0022 (2911)
26  forest → category0023 (4457)
27  freeway → category0024 (800)
```

```
gamespace → category0025 (1600)
golfcourse → category0026 (800)
grave → category0027 (465)
ground_track_field → category0028 (700)
harbor → category0029 (2117)
highway → category0030 (306)
hirst → category0031 (1046)
industrial_area → category0032 (700)
intersection → category0033 (1320)
island → category0034 (700)
lake → category0035 (700)
mangrove → category0036 (963)
meadow → category0037 (700)
mediumresidential → category0038 (800)
mobilehomepark → category0039 (100)
mountain → category0040 (1664)
mountain_road → category0041 (1066)
overpass → category0042 (1918)
palace → category0043 (700)
parkinglot → category0044 (2799)
pipeline → category0045 (353)
railway → category0046 (1116)
railway_station → category0047 (700)
rectangular_farmland → category0048 (700)
river → category0049 (2071)
roundabout → category0050 (700)
runway → category0051 (1271)
sea → category0052 (1494)
snow_mountain → category0053 (967)
sparseresidential → category0054 (1600)
stadium → category0055 (700)
storagetanks → category0056 (2044)
stream → category0057 (1109)
tennis_court → category0058 (800)
terrace → category0059 (700)
thermal_power_station → category0060 (700)
tower → category0061 (173)
turning_circle → category0062 (309)
wetland → category0063 (700)

Total images in step2 finetuning: 67801
```

The anonymized processing of dataset in step 2 is similar in step 1 where we perform a **global category mapping** that deterministically converts all original semantic classes into **anonymous category IDs** (e.g., `airfield → category0001`, ⋯, `wetland → category0063`), and

the dataset is rebuilt using only these mapped IDs. After this preprocessing, the training data **no longer contains any original class names.**

S2.4 Second-Step Finetuning

We apply the following prompt structure as our second-step finetuning period.

Second-Step Finetuning Prompt

```
def build_instruction() -> str:
    return (
        "You are a remote-sensing imagery expert. <image>\n\n"
        "You are provided with an image accompanied by detailed segmentation results generated by the SAM model. "
        "Each segment includes:\n"
        "- 'bbox': bounding box coordinates [x,y,w,h], approximately locating the segmented object;\n"
        "- 'area': the area (in pixels) of the segmented object;\n"
        "- 'counts': pixel-level RLE (Run-Length Encoding) mask precisely defining object boundaries;\n"
        "- 'score': confidence score assigned by SAM (optional).\n\n"

        "STRICTLY follow these INTERNAL THINKING STEPS (do NOT output your thinking process, these steps are for internal reasoning ONLY):\n\n"

        "Step 1: INTERNALLY analyze the image comprehensively by observing its overall textures, patterns, and visual structure. "
        "Carefully consider the provided segmentation results (bbox, area, pixel-level masks) to identify pixel-level, object-level and region-level features.\n\n"

        "Step 2: Based on your analysis, INTERNALLY select the MOST APPROPRIATE first-level land-use category from the following standardized list:\n"
        "[Cultivated Land, Garden Land, Forest Land, Grassland, Commercial Service Land, Industrial, Mining, and Storage Land, Residential Land, Public Administration and Public Service Land, Special Land, Transportation Land, Water Bodies and Hydraulic Facility Land, Other Land].\n\n"

        "Step 3: INTERNALLY determine the DETAILED subtype within the selected first-level land-use category, precisely guided by the segmentation information provided by SAM.\n\n"

        "Step 4: INTERNALLY form a clear and precise description of this image, explicitly leveraging the SAM segmentation results (bbox locations,
```

```
    mask shapes, areas, and pixel-level information) to consolidate your
    understanding and classification decision.\n\n"
22
23          "IMPORTANT OUTPUT REQUIREMENT:\n"
24          "You MUST ONLY OUTPUT the final classification category label
    corresponding to the detailed land-use subtype (e.g., category0011). "
25          "DO NOT OUTPUT any intermediate reasoning, first-level categories,
    subtype names, or image descriptions. "
26          "ONLY OUTPUT a single category ID."
27      )
28
```

The first-level land-use list in the prompt is **not used as label supervision** and does **not introduce semantic leakage**. Our dataset and training targets have already been anonymized to **ID-only category codes**, and the model is explicitly constrained to output **only a single anonymous category ID** (e.g., `category0011`), never any textual class name.

The first-level list is included solely as a **fixed reasoning scaffold** to enforce a **standardized and consistent decision process**, helping the model converge to a unified, deterministic output format under an ID-only objective. Therefore, the learning signal remains strictly **image/segmentation → anonymous ID**, fully decoupled from textual label semantics.

In Step-II training, each sample consists of an image tile paired with the serialized geometric metadata of **all** SAM2-predicted regions in that tile, including bounding boxes, areas, and run-length encoded (RLE) masks.
While the supervised target is an anonymized category ID, **the model performs taxonomy-grounded interpretation conditioned on the collection of region masks**, using them as explicit spatial evidence rather than treating the image as an unstructured global scene.

S3: LLM-Judge Evaluation

S3.1 Prompt

S3.1.1 Naturalness

Role Definition

You are an impartial writing evaluator. Your job is to judge how natural and human-like each DESCRIPTION is.

Scope & Constraints

- Work ONLY with the given text, do not use outside knowledge and no fact-checking.

Input

INPUT (DESCRIPTION):

<<DATASET_JSON>>

Output Requirements (STRICT)

- Return exactly five numbers in this order, separated by single spaces:
  a. Grammar & Syntax
  b. Lexical Naturalness & Idiomaticity
  c. Discourse Coherence & Flow
  d. Style & Register Appropriateness
  e. Human-likeness vs Machine "Tells"
- Evaluate in the natural language of the DESCRIPTION.
- For code-mixed text, It is not a coherent and logical natural language; if unclear, reflect this in Coherence and Human-likeness.
- Fragments, lists, bullet points, or telegraphic style can be appropriate: assess "Style & Register" relative to whether that style fits a typical description.
- Penalize machine "tells": templated phrasing, repetitive n-grams, self-disclosure (e.g., "As an AI…"), over-hedging, rigid slot-filling, or contradictions.
- Score each DESCRIPTION independently; do not normalize across items.

Scoring Rules

- Each sub-score must be within [0, 5] using 0.5 increments only (allowed set: 0, 0.5, 1, 1.5, ⋯, 5).
- Round to the nearest 0.5 when needed.

Formatting Rules

- Use a dot as the decimal separator.
- No commas, brackets, units, markdown, explanations, or any extra text.
- Output must be a single line with single spaces, no leading/trailing spaces.no markdown, no explanations, no extra text.

Special Case

- If the DESCRIPTION is empty or non-textual gibberish, output:
  0 0 0 0 0

S3.1.2 Informative

Role Definition

You are an impartial writing evaluator. Rate how detailed, specific, and informative each DESCRIPTION is.

Scope & Constraints

- Work ONLY with the given text, do not use outside knowledge and no fact-checking.

Input

INPUT (DESCRIPTION):

<<DATASET_JSON>>

Output Requirements (STRICT)

- Return exactly five numbers in this order, separated by single spaces:
  a. Coverage of Key Facets — Breadth of major elements relevant to the described item/topic (within the text itself).
  b. Specificity & Quantification — Presence of precise facts (numbers, named entities, model/versions), measurable details, and exact qualifiers.
  c. Concreteness & Observability — Tangible attributes a reader could verify from the text alone (actions, measurements, materials, examples).
  d. Context, Constraints & Relations — Purpose, conditions, assumptions, dependencies, comparisons, trade-offs, timelines.
  e. Relevance & Non-Redundancy — Focus on the topic without filler/boilerplate; minimal repetition of the same points.

Scoring Rules

- Each sub-score must be within [0, 5] using 0.5 increments only (allowed set: 0, 0.5, 1, 1.5, ⋯, 5).
- Round to the nearest 0.5 when needed.

Formatting Rules

- Use a dot as the decimal separator.
- No commas, brackets, units, markdown, explanations, or any extra text.
- Output must be a single line with single spaces, no leading/trailing spaces.no markdown, no explanations, no extra text.

Special Case

- If the DESCRIPTION is empty or non-textual gibberish, output:
  0 0 0 0 0

S3.2 LLM-Scoring Rule

We also describe the scoring rules in detail in the main text; here we provide a more fine-grained scoring rubric:

Scoring Rules

For each region-level description, we use GPT-4o as a rubric-based judge to score two axes: Naturalness and Informativeness. Each axis contains five weighted sub-modules (as listed in Table S3). For every sub-module, the judge outputs a score from 0 to 5 using 0.5 increments only (with 0 the worst and 5 the best)

Score aggregation

For each axis, we compute a weighted score and scale it to [0, 100]: we first normalize each sub-score by dividing by 5, multiply by its weight, sum over the five sub-modules (weights sum to 1), and finally multiply by 100. We report Naturalness and Informativeness separately; when a single overall scalar is needed, we use the average of the two axis scores.

Determinism and formatting

To reduce randomness and ensure reproducible outputs, we set the sampling temperature to 0 and enforce a single-line fixed numeric output format via structured prompting.

Dimension	Sub-module	Weight
Naturalness	Grammar & Syntax (G)	0.25
	Discourse Coherence & Flow (D)	0.25
	Lexical Naturalness & Idiomaticity (L)	0.20
	Style & Register Appropriateness (S)	0.15
	Human-likeness vs Machine "Tells" (H)	0.15
Informativeness	Coverage of Key Facets (C)	0.25
	Specificity & Quantification (S)	0.25
	Concreteness & Observability (O)	0.20
	Context, Constraints & Relations (X)	0.20
	Relevance & Non-Redundancy (R)	0.10

S3.3 LLM-as-Judge Calibration, Reliablity, Abnormal-Case Analysis and Rules

We employ GPT-4o as an LLM-as-judge to produce scalable, rubric-based scores under a predefined scoring guideline and a strict output schema. All samples are evaluated using identical criteria and the same automatic parsing rules to ensure comparability across methods.

Reliability and stability of judge scoring. We evaluate the stability of the LLM-as-judge by checking score robustness under repeated scoring of the same inputs (e.g., measuring score variance and agreement rate across multiple runs). This helps verify that the judge provides sufficiently consistent signals for comparative evaluation.

Zero-score policy for irregular outputs. To prevent unparseable or non-compliant responses from introducing bias, we assign a score of 0 to any irregular output, including but not limited to:

· Empty outputs (empty string, whitespace-only responses, or missing valid content);

· Format violations (missing required fields, not parsable into the predefined schema, or structural corruption);

· Clearly invalid content (placeholders, repetitive nonsensical text, or content unrelated to the input).

Handling overlong/truncated outputs. For outputs that are truncated due to length limits, we follow a “score what is observable” principle: the judge scores the response based on the available (untruncated) portion. If critical information is missing because it appears after the truncation point, the score is reduced accordingly; truncation itself does not automatically trigger a zero score.

Abnormal-case diagnosis. We further analyze zero-score or abnormal cases and summarize the major failure modes, such as empty outputs, format violations, truncation-induced incompleteness, and off-topic/unsupported content. This diagnosis improves interpretability and helps identify systematic issues in generation and evaluation.

S4: Manual Evaluation

S4.1 Manual Evaluation Scoring Rule

We also describe the scoring rules in detail in the main text; here we provide a more fine-grained scoring rubric:

Ground-truth Level-1/Level-2 labels for manual scoring are produced via expert re-annotation rather than directly taken from LoveDA names. Specifically, we randomly select 449 LoveDA tiles and evaluate region-text pairs: each “sample” corresponds to one predicted mask/region and its generated tag/description. Five remote-sensing/GIS researchers independently assign the reference Level-1 and Level-2 category for each region following the Chinese Land Use

Classification Category (Reference 17 and Appendix S1) by inspecting the RGB tile with the mask overlay (and the region’ s dominant pixels). For descriptions, annotators judge whether the generated text correctly characterizes the same region’ s land-use/land-cover attributes and context, using the image–mask evidence as reference (0 = incorrect/mismatched; 0.5 = partially correct but incomplete or mixed; 1 = fully correct and consistent). Disagreements are resolved by majority vote with final adjudication; predicted tags are scored by exact match (0/1) against the consensus Level-1/Level-2 reference, and description scores are averaged across annotators.

Scoring Item	**Score Rule**
Level-1 Category	0 / 1
Level-2 Category	0 / 1
Description	0 / 0.5 / 1
Scoring Logic	
Level-1 / Level-2 incorrect	0
Description partly correct	0.5
Description fully correct	1
Total Sample Score: 3 points	

S4.2 Manual Evaluation Protocol, Annotator Details, Fairness Measures and Rules

We complement automatic evaluation with a human assessment protocol designed to be fair, consistent, and domain-appropriate. We recruit five annotators with substantial experience in remote sensing and related fields to score model outputs

Training and guideline standardization. Before formal evaluation, all annotators undergo training to ensure consistent interpretation of the rubric, including walkthroughs of scoring criteria, calibration with representative examples, and discussion of borderline cases. Annotators follow a unified written guideline during evaluation.

Scoring and aggregation. Each sample is scored independently by the annotators according to the same rubric. The final human score is obtained by aggregating individual ratings (e.g., mean across annotators), and disagreements are handled according to the guideline to ensure consistency.

Fairness control. To maintain fairness, all methods are evaluated under identical conditions and with the same scoring criteria. Annotators are instructed to base their judgments strictly on task-relevant quality and compliance, and to avoid using stylistic preferences or non-essential cues as scoring factors. The training phase and calibration rounds are used to reduce inter-annotator variance and improve evaluation reliability.

All five annotators were trained with unified guidelines and calibration examples to align scoring criteria. We further report the inter-annotator agreement for Level-1/Level-2 labeling and the ordinal description scores, showing substantial consistency across annotators and supporting the reliability of the manual evaluation.

S5: Detailed Setting Illustration in Stage I: SAM2 Segmentation

In Stage I, we adopt SAM2 as a class-agnostic mask proposer, using the official checkpoint as the off-the-shelf baseline and further fine-tuning it on the full OpenEarthMap to obtain an adapted variant. For evaluation on LoveDA (where urban scenes often contain highly fragmented regions and thin, complex structures), we derive instance proxies by decomposing each semantic class mask into per-class connected components, and match predicted masks to these GT instances via one-to-one IoU assignment (unmatched GT = 0) to compute instance mIoU/mDice.