APERTURE: TRAINING-FREE MULTISCALE CONCEPT BOTTLENECKS FOR REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

While earth observation models have advanced substantially, they still lack interpretability. While concept-bottleneck models provide interpretability and expert interaction, they are either too expensive to train for the remote sensing domain or perform poorly without annotation. We posit that in expert domains like remote sensing, such training-free models require both fine details in both image and concept space. In image space, we propose a multiscale concept bottleneck using greedy quadtree routing to locate small concepts. In concept space, we replace contrastive vision language models with pre-trained MLLMs and present a way to get reliable concept scores from them. We introduce APERTURE that blends concept scores at the global image and native concept-scale level to give state-of-the-art training-free model performance. To test these models, introduce SiFC, a fine-grained concept-centric dataset across three countries, with human-reviewed class-level concept maps. On SiFC, APERTURE outperforms the best training-free baselines by more than 10 percentage points in macro F1-score, and notably also outperforms supervised concept bottleneck models. Targeted component-removal tests examine whether concept scores respond to changes in visual evidence, while temporal experiments show that descriptor updates improve recognition of technological changes without retraining. Our data and code are available at `https://anonymous.4open.science/r/aperture-C1984/`.

1 INTRODUCTION

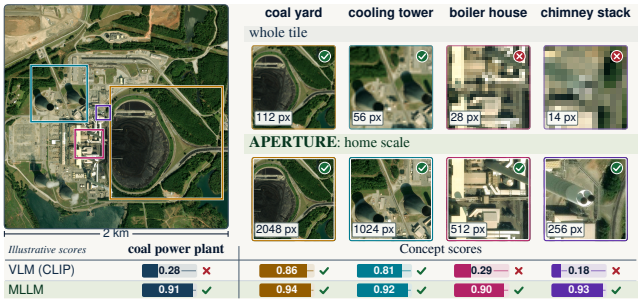


Figure 1: **Multi-scale concept bottleneck:** Resizing the 4096-pixel coal plant (spanning 4km^2) shrinks the coal yard and chimney to 112 and 14 pixels (top right). APERTURE instead crops concepts at home scales, enabling it to recognize both parts and correctly classify the site, unlike CLIP.

Earth observation enables continuous monitoring of various events and concepts, such as deforestation or the construction of new industries, so that rapid actions can be taken to protect our environment. Visual recognition in satellite imagery has advanced substantially over the last decade, driven by progress in fine-grained (Chu et al., 2024), self-supervised (Ayush et al., 2021; Cong et al., 2022), vision-language (Li et al., 2023; Hu et al., 2025), and reasoning models (Yao et al., 2026; Liu et al., 2026). Yet, interpretable models remain underexplored. Remote sensing is an expert domain that requires close interaction between model users (domain experts) and models. Interpretable models, especially ante-hoc (or interpretable-by-design) models, are specifically useful in such cases, as the experts can not only understand the model's decision but can also intervene and improve them.

Concept bottleneck models (CBMs) (Koh et al., 2020) are one way to create such ante-hoc models. CBMs make class recognition decisions on top of an intermediate list of human-understandable concepts, making the decision more transparent and intervenable. However, they require not only

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

class-level but also expensive concept-level annotations, making data collection significantly time-consuming (Mall et al., 2021). As a result, several follow-up works have proposed both concept label-free (Oikarinen et al., 2023) and completely training-free variants Menon & Vondrick (2023).

Training-free CBMs are ideal for expert domains such as ecology or remote sensing, as often we do not have extensive labeled training data, and novel classes can arise at deployment time. They allow for easy class and concept addition at deployment and are therefore ideal for this scenario. However, existing training-free CBM methods still struggle when applied to satellite image recognition.

Firstly, training-free CBMs rely on contrastive vision-language models such as CLIP (Radford et al., 2021) to recognize concepts. But these models struggle with non-object-centric attribute recognition (Kang et al., 2025), especially in fine-grained expert domains such as remote sensing. We address this issue by replacing contrastive VLMs with a multimodal large-language model (MLLM)-based formulation. This enables us to reason better about domain-specific concepts and attributes.

Using an MLLM inside a CBM introduces a second, non-trivial problem. Generative MLLMs naturally return discrete presence/absence decisions, whereas CBMs rely on continuous concept scores. Asking an MLLM to further explicitly report its confidence results in unreliable and miscalibrated scores. Borrowing from binary question answering works (Giovannotti & Gammerman, 2024), we derive continuous concept evidence from the normalized next-token probabilities assigned to `Yes` and `No`. This yields significantly more effective concept scores. To our knowledge, we are the first to apply MLLMs to training-free CBMs and use the decoding probability to score concept confidence.

Satellite imagery introduces a third challenge: relevant concepts may occupy vastly different spatial extents. For example, as shown in fig. 1, coal power plant may span $4km^2$, but an important concept, such as a chimney stack, only occupies $0.016km^2$ ($< 0.4\%$ of the image). This makes recognition of these small concepts very difficult. Therefore, we introduce a multi-scale concept recognition framework. Instead of just looking for a concept at the global image scale, we recursively break down images into smaller scales to search and aggregate concept scores across them. However, exhaustively evaluating every crop is computationally prohibitive. Recursively splitting an image into 4 smaller quadrants 4 times (see fig. 1), requires evaluating $341\times$ more regions than global evaluation. Therefore, we introduce a greedy quadtree search that recursively searches within only one crop at every zoom level, reducing the search time from exponential to linear in quadtree depth.

More importantly, satellite imagery provides information typically unavailable in ordinary internet imagery: *known ground sampling distance (GSD)*. We exploit this information to associate each concept with a characteristic physical "home scale", allowing the search to stop at the resolution at which the concept should be naturally recognized.

These limitations are difficult to study with the existing remote-sensing benchmarks, which mostly contain atomic scene categories (Helber et al., 2019; Sumbul et al., 2019), recognizable without identifying constituent parts and attributes. We therefore introduce a new dataset, SiFC (Satellite image Facility Concepts), with 800 images from 6 fine-grained industrial classes. We additionally construct a subset of FMoW (Christie et al., 2018), consisting of similar multi-concept fine-grained categories. Both these datasets span multiple world regions to test out-of-region generalization.

Together, these components turn training-free CBMs from a global contrastive-similarity model into a generative, spatially adaptive, and physically grounded concept recognition framework for satellite imagery that we call APERTURE. We show that APERTURE achieves state-of-the-art performance among prior training-free CBMs on both SiFC and FMoW. Notably, APERTURE even outperforms supervised CBMs on out-of-region images. We also show how our training-free method can easily allow interventions from experts when a class concept changes over time. Our contributions are:

- We introduce an MLLM-based training-free concept bottleneck for fine-grained satellite recognition and derive continuous concept evidence from binary next-token probabilities for it.
- We introduce efficient greedy quadtree routing for discovering concepts that may occupy only a small fraction of the full image.
- We introduce *home scale*, exploiting known GSD in satellite images to associate concepts with their characteristic physical extent and constrain multiscale recognition to an appropriate resolution.
- We introduce a concept-centric benchmark, SiFC, where we achieve state-of-the-art training-free performance, even outperforming supervised interpretable models under geographic shifts

2 Related Works

Interpretable recognition: Post-hoc methods, such as Grad-CAM (Selvaraju et al., 2017; Chattopadhay et al., 2018), explain predictions without guaranteeing faithful insight into the model's decision process (Adebayo et al., 2018; Rudin, 2019). This motivated ante-hoc models that incorporate human-readable concepts into the prediction process. (Chen et al., 2019). Among ante-hoc methods, Concept Bottleneck Models (CBMs) (Koh et al., 2020; Espinosa Zarlenga et al., 2022; Kim et al., 2023; Espinosa Zarlenga et al., 2023) have become a prominent approach in expert domains, as they provide an intermediate human-interpretable concept layer. However, standard CBMs require both class labels and image-level concept annotations, the latter of which are significantly expensive to collect at scale (Mall et al., 2021). Label-free CBMs remove concept annotations through pretrained vision-language models but still require labeled images to learn the class predictor (Oikarinen et al., 2023; Yang et al., 2023; Srivastava et al., 2024). More recently, entirely training-free CBMs, such as classification-by-description (Menon & Vondrick, 2023; Pratt et al., 2023; Chen et al., 2025; Yamaguchi et al., 2025), have been introduced. They remove target-specific fitting entirely and require neither labeled images nor image-level concept annotations. We show that existing training-free CBMs struggle with fine-grained interpretable recognition in satellite imagery and introduce a new framework for it. While our work explores a spatial hierarchy over image regions, prior works have explored semantic hierarchies among concepts (Pittino et al., 2023; Hill et al., 2026).

Interpretable recognition in satellite imagery: Concept-based explanations of satellite scene classification aim to link class decisions to human-interpretable visual concepts. Existing methods highlight influential regions (Kakogeorgiou & Karantzalos, 2021) or generate textual reasoning before answering visual questions (Wang et al., 2024). However, these approaches do not explicitly expose how intermediate concepts contribute to the prediction. PCINet (Hua et al., 2024) is related to our setting, using LLM-generated concepts for interpretable aerial-image recognition. However, it evaluates on MAI dataset (Hua et al., 2021), a scene-recognition benchmark with broad categories. In contrast, SiFC studies fine-grained compositional recognition, where competing system-level facility classes share constituent parts and attributes, and recognition therefore requires distinguishing classes through their combinations. Furthermore, SiFC is constructed with the same classes represented across three geographic regions, enabling evaluation of out-of-region generalization.

3 Satellite Imagery Facility Concepts (SiFC)

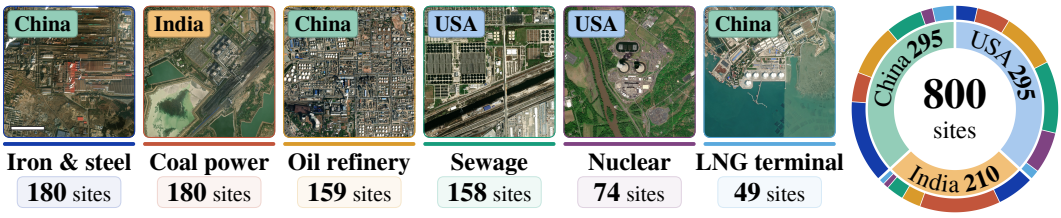


Figure 2: **SiFC at a glance.** (Left) sample facility classes, labeled with their country and the number of sites. Each tile covers about $4km^2$ (4096×4096 px at zoom level 18, World Imagery Wayback (Esri)); the white bar is 500 m. (Right) 800 sites by country (inner) and classes (outer).

Existing recognition datasets for remote sensing largely focus on very coarse scene or land-cover categories, which do require limited reasoning about parts and their attributes. For example, widely used datasets such as EuroSAT (Helber et al., 2019) and BigEarthNet (Sumbul et al., 2019) primarily contain land-cover categories that are not naturally characterized through a composition of semantically meaningful concepts. Even higher-resolution scene-recognition datasets such as RESISC45 (Cheng et al., 2017) include many relatively atomic categories, such as baseball fields. In contrast, we seek *system*-like entities that can be characterized by the presence, arrangement, and attributes of their constituent components. Industrial facilities naturally fit this setting: steel and power plants contain recurring functional sub-components that support recognition by parts and attributes.

To support this setting, we introduce **Satellite Imagery Facility Concepts (SiFC)**, a dataset for concept-based classification of industrial facilities in satellite imagery. SiFC contains 800 samples

from *three regions*: USA, India, and China and covers *six facility* classes. This geographic diversity captures variation in facility appearance, surrounding land use, and component layout Images are sourced from the public ESRI World Imagery Wayback archive. We release corner coordinates with retrieval scripts rather than redistribute imagery. Each image has a resolution of 4096×4096 pixels and is retrieved at zoom level 18, covering an area of approximately $4\,\mathrm{km}^2$. Samples include facility and country labels, while each class is paired with a human-annotated concept map and country-specific component descriptions and layout. SiFC supports direct zero-shot classification, training-free, and concept label-free CBM. It also supports testing out-of-region generalization.

4 METHOD

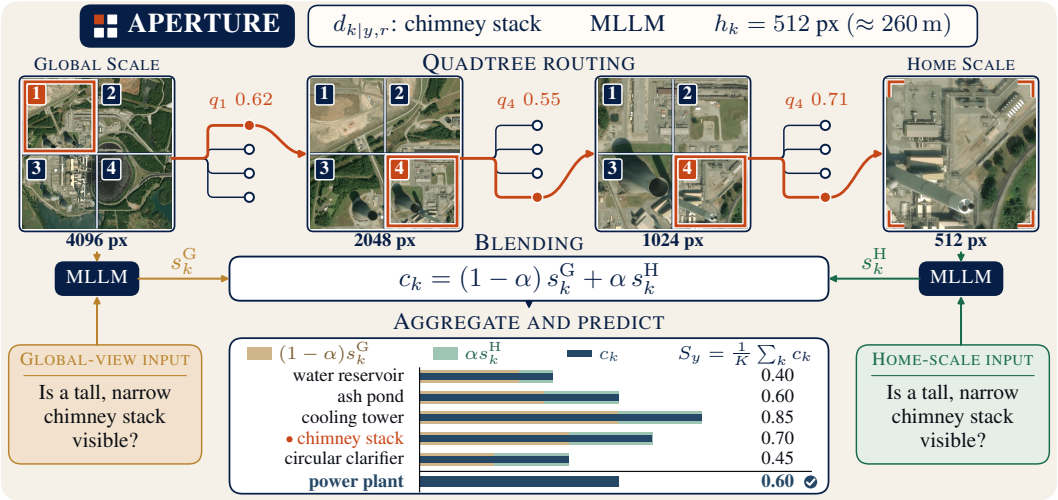


Figure 3: Overview of APERTURE. For each class- and country-conditioned descriptor $d_{k|y,r}$, a frozen MLLM follows one quadtree branch to its home scale h_k while pruning the remaining quadrants. Global and home-scale answer scores are blended into the concept score c_k^{B}. The concept scores are then averaged into the class score S_y, and the highest-scoring class is predicted.

We aim to build an interpretable-by-design recognition model for remote sensing categories. In section 4.1, we first introduce variants of CBMs (Koh et al., 2020), as it forms the basis of our ante-hoc interpretable recognition method. We use training-free (CBMs) as a starting point, and in sections 4.3 to 4.5, we introduce our contributions to improve training-free CBMs.

4.1 BACKGROUND

Concept Bottleneck Model: CBMs (Koh et al., 2020) predict classes through an intermediate vector of interpretable concepts, $\hat{\mathbf{c}} = g(x)$, followed by a task predictor $\hat{y} = f(\hat{\mathbf{c}})$. This intermediate representation enables interpretability and concept-level interventions. However, supervised CBMs require costly concept-annotated triplets $\{x^{(i)}, \mathbf{c}^{(i)}, y^{(i)}\}_{i=1}^{N}$, especially in expert domains.

CBMs with pre-trained VLM: Recent methods (Oikarinen et al., 2023) use LLMs (Brown et al., 2020) or large concept dictionaries (Oikarinen et al., 2023), to construct a concept set $\mathcal{C}$ and to reduce annotation cost. Pre-trained VLMs like CLIP models g, and image-class pairs are used $\{(x^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$ to learn f. No expensive manual concept annotations are needed, but in many cases they learn incorrect correlations between classes and concepts that are not human-aligned. In the following discussion we call these methods Concept Label-free CBMs (CLF-CBMs).

Training-free CBMs: Training-free CBMs remove target-specific supervision entirely. They require no labeled target images, no concept annotations, and no parameter optimization on the target dataset. Like CLF-CBMs, LLMs construct a list of concepts; however, along with the list, the LLMs also provide a mapping between classes and specific concepts (Menon & Vondrick, 2023). As a result, the image-concept function g, is replaced with a pre-trained VLM, and the

concept-class function f gets information from an LLM. For each class i, the LLM selects a subset $\mathcal{C}_i \subseteq \mathcal{C}$ from the complete concept set. Given an image's concept vector $\hat{\mathbf{c}}$, the class score is $\hat{y}_i = f_i(\hat{\mathbf{c}}) = \frac{1}{|\mathcal{C}_i|} \sum_{k \in \mathcal{C}_i} \hat{c}_k$, where $\hat{c}_k$ is the score for concept k.

We choose to build our method from TF-CBMs. While TF-CBMs result in weaker-performing models compared to standard and CLF-CBMs, they remove expensive concept-label collection and target-specific training and are preferred in many cases, like ours, where data collection is expensive. Moreover, in our results we observe that TF-CBMs also limit overfitting by avoiding learning spurious correlations. For our problem of satellite image classification, this is especially useful, as we show that TF-CBM models generalize much better to out-of-region images compared to LF-CBMs. Furthermore, TF-CBM allows intervention and correction at the concept level without retraining.

4.2 Problem Setup and Overview

Given a satellite image x and country metadata r, we predict its facility class $y \in \mathcal{Y}$ without labeled target images, image-level concept annotations, or target-data training. We build on classification-by-description (CbD) (Menon & Vondrick, 2023), a TF-CBM. It represents each class using a set of concepts/descriptors, with the prepended class name, and uses CLIP as g.

However, applying CbD to satellite imagery presents two challenges. While CLIP and similar contrastive vision-language models understand global image concepts well, they struggle to understand part and local attributes (Kang et al., 2025). This problem is further exacerbated when the concepts belong to expert domains like satellite images. Another challenge is that industrial facilities span large areas and contain components at widely varying scales, so passing the full image to the vision encoder may obscure small but discriminative structures (e.g., a storage tank can be recognized at a coarse facility scale, whereas thin flare stacks require a finer local scale).

These challenges motivate us to revise how concepts are recognized. To tackle the first challenge, we replace CLIP-like VLMs with general multimodal large-language models (MLLM). We discuss the details of these modifications in section 4.3. Replacing contrastive VLMs with MLLMs also requires updating the concept definitions; we discuss them in section 4.4. To tackle the second challenge, we present a methodology to perform concept recognition at multiple scales in section 4.5

4.3 From CLIP to MLLM

CbD scores each concept with a zero-shot CLIP-style encoder, but it does not work well on satellite images. CLIP (Radford et al., 2021) itself reports that it is "quite weak on satellite image classification (EuroSat (Helber et al., 2019) and RESISC45 (Cheng et al., 2017))". This issue is further amplified when CLIP is used for recognizing localized concepts instead of global classes. Furthermore, since remote-sensing VLMs (Mall et al., 2024; Liu et al., 2024) are trained with similar methodology, they also fail to recognize concepts, even though they can recognize classes.

Since CbD relies on this alignment to score descriptors, it inherits the same limitation. We therefore replace CLIP with an MLLM. We choose general-purpose MLLMs (e.g., Gemma (Team et al., 2026)), instead of remote sensing MLLMs, again, as remote sensing MLLMs are better at understanding classes in satellite images, but general-purpose MLLMs are much better at concepts. To check for the presence of a concept (say, "thin flare stacks"), instead of measuring the cosine similarity via CLIP, we ask an MLLM: `In the image is a thin flare stack present?`

From binary answers to continuous concept scoring: However, when replacing a contrastive VLM with an MLLM, one challenge is that we do not get continuous scores and instead get binary answers. CbD and CBM inputs require continuous concept scores (Koh et al., 2020; Espinosa Zarlenga et al., 2022; Menon & Vondrick, 2023) and binarizing concept scores results in poorer performance, as it results in frequent ties and treats weak and strong evidence equally.

One solution is to ask the model to report a confidence value with its answer (`e.g., ''yes, with 90% confidence''`). However, we find that the verbalized confidence is unreliable because LLMs and VLMs are overconfident and poorly calibrated (Xiong et al., 2024). Across 35,200 answers, Gemma-4-31B assigned 98.7% of them a score of at least 0.9 and produced similar mean confidence for present and absent answers, 0.934 and 0.933, respectively. These scores provide less variance, and as a result, the model behaves like a binary decision-based CBM (Table 1).

We instead present an alternative strategy to obtain continuous scores from MLLMs. We compute the raw decoding logit scores when answering `yes` and `no`. For each descriptor d_k, we construct the binary question "`Is` d_k `visible in this satellite image? Answer yes or no.`" We obtain the logits for `yes` (z_1) and `no` (z_0), which we convert into a continuous concept score using the temperature-scaled softmax $p(\texttt{yes} \mid x) = \exp(z_1/\tau) / [\exp(z_1/\tau) + \exp(z_0/\tau)]$. Compared with verbalized confidence, binary-answer probabilities provide a better-calibrated estimate of model uncertainty. Although decoding probabilities have previously been used to measure model uncertainty (Kadavath et al., 2022), we are the first to use them in concept bottlenecks.

4.4 Changes to the concept Definitions

Contrastive VLMs prefer shorter text queries (limited to 64 tokens for CLIP); so LLM-generated descriptors are typically intentionally kept shorter. However, MLLMs support longer contexts and better grounding of sentences and paragraphs. Therefore, we adapt our prompting as follows. For each of these modifications, more details and exact prompts are present in appendix A.1.

From Attribute Phrases to Structured Descriptors. We observe that MLLMs respond more reliably to complete sentences with explicit visual cues (table 1). We therefore expand each descriptor with additional details about *appearance, spatial placement, and likely confusers.* This structure provides both supporting and contrasting evidence, making concept verification less ambiguous.

From Explicit Class Names to Class- and Country-Conditioned Descriptors. The original CbD prepends class names to the descriptors, improving context and thus performance but hurting interpretability. In our case, prepending labels not only limits interpretability, it also reduces performance. Therefore, instead of prepending the class name, we prompt the LLM to modify the description conditioned on the class name. Facilities of the same class can differ across countries; therefore, for generalization, we can also condition descriptions on country-specific information.

Concept-Specific Home-Scale Assignment. Unlike internet images, objects in satellite imagery often have a well-defined physical scale. We exploit this *home scale* to search for concepts at an appropriate scale. The descriptor-generation LLM assigns each descriptor a home scale $h_k \in \mathcal{S}$, defining the crop size that balances concept visibility and recognition context.

4.5 Concept-Dependent Spatial Scale.

Satellite image concepts could cover different ground areas. Large concepts require a wide view to preserve their shape and layout. Small but important concepts require a local "zoom-in" for a local view to remain visible. This object scale variation is a known challenge in aerial imagery (Xia et al., 2018), and prior work shows that remote-sensing representations benefit from explicit scale information (Reed et al., 2023; Revankar et al., 2025). We observe when a large $4\,\mathrm{km}^2$ tile is resized to the input resolution, small concepts are difficult to resolve. An LNG storage sphere (small concept) may occupy only two or three pixels, making it harder for the MLLM to recognize (Khayatkhoei et al., 2025; Wang et al., 2025). Therefore, we do detection at multiple scales.

One option is to zoom in to each concept's home scale, score every crop along the way, and keep the highest score at each level. We use crop widths in $\mathcal{S} = \{4096, 2048, 1024, 512, 256\}$ pixels. Starting from the full image, we stop zooming when we reach the concept's home scale $h_k \in \mathcal{S}$. Scoring every crop is costly, as crops increase exponentially with every zoom level. We therefore use a greedy *quadtree search*; at each level, the MLLM selects the quadrant most likely to contain the concept. We zoom into that quadrant until we reach h_k. Each step halves the crop width (e.g., $4096 \to 2048 \to 1024$ pixels), so we search only one path through the quadtree. MLLM-based routing decides where to look, but we do not use its concept score at intermediate levels. At h_k, we ask whether the concept is visible in the selected crop and use $p(\texttt{yes})$. With $\ell_k = \log_2(4096/h_k)$ zoom levels, exhaustive search requires $O(4^{\ell_k})$ MLLM calls, while greedy routing requires $O(\ell_k + 1)$; at $h_k = 256$, this is 341 versus 5 calls.

The global view preserves facility layout, while the home-scale crop reveals small components. Therefore, we perform a linear blending of these global: s_k^{G} and home-scale: s_k^{H} concept scores to get a final concept score: $c_k = (1-\alpha)s_k^{\mathrm{G}} + \alpha s_k^{\mathrm{H}}$. Here, $\alpha \in [0, 1]$ controls the home-scale contribution. Like CbD, the class score is the average of individual *blended* concept scores.

5 RESULTS

5.1 IMPLEMENTATION DETAILS AND BASELINES

We primarily evaluate APERTURE on SiFC and the fMoW facility subset. Across tasks we report the macro-F1 (%) score averaged over classes, both per-country and by pooling all images denoted by "All". Baselines with supervised training are trained on a leave-one-country-out setup.

Global or home-scale views are resized to 224×224 images, matching CLIP ViT-L/14. Home-scale crops are extracted before resizing. Following prior binary QA work (Giovannotti & Gammerman, 2024), the temperature for question answering is set to $\tau = 30$. Global and home-scale scores are blended with $\alpha = 0.25$. Our sensitivity analysis in the appendix shows that α values between $[0.2, 0.5]$, result in similar model performance. Unless specified, Gemma-4-31B (Team et al., 2026)is our concept recognition backbone, but we also test other backbones in the ablation.

We compare with several training-free CBM baselines: zero-shot CLIP (Radford et al., 2021), CbD (Menon & Vondrick, 2023), and LaZSL (Chen et al., 2025), and supervised concept label-free CBMs: CF-CBM (Panousis et al., 2024), LaBo (Yang et al., 2023), and VLG-CBM (Srivastava et al., 2024). We also compare two MLLM concept-scoring strategies: *conf.*, which asks MLLM for a confidence score and *prob.,* is our method using the next-token probability. We also compare separate global and home-scale concept predictions and their blend (APERTURE) and also compare prompting strategies ranging from general to class-country-conditioned descriptors.

5.2 DISCUSSION

Table 1: Macro-F1 (%; ↑) on SiFC and the fMoW facility subset. The highest and second-highest scores in each column across all rows are **bold** and underlined, respectively. All Gemma settings use the same frozen MLLM. TF denotes no task-specific training.

	Model	Setting	TF	Descriptor	Score	SiFC				fMoW facility			
						India	USA	China	**All**	USA	France	Russia	**All**
Baselines		CLIP zero-shot	✓	CND	cos.	53.03	67.19	53.29	62.16	71.29	57.71	51.91	65.85
		CBD	✓	CBD	cos.	59.37	66.60	57.73	66.42	70.57	54.11	51.02	62.74
		LaZSL	✓	AD	OT	69.38	72.83	58.78	69.76	**78.63**	59.86	57.17	69.90
		CF-CBM (high)	✗	HCD	cos.	40.25	25.67	48.60	39.35	12.60	25.89	24.11	23.18
		CF-CBM (low)	✗	HCD	cos.	53.92	40.87	60.69	49.88	41.30	72.52	58.17	55.92
		LaBo	✗	LC	cos.	66.70	62.97	53.39	60.44	64.75	57.50	65.21	65.11
		VLG-CBM	✗	LC	logit	71.91	64.04	66.88	68.21	69.40	68.26	62.28	71.62
Gemma 4 31B		Class name only	✓	CND	–	60.62	68.02	62.89	70.40	67.52	74.98	61.46	73.18
	Our TF Setting	Home scale	✓	HSR-CCCD	prob.	41.74	66.10	45.56	55.09	57.76	77.01	49.04	63.46
		Global Scale	✓	RAG-CCCD	prob.	70.68	75.54	74.67	77.53	69.69	75.17	70.07	74.40
			✓	CCCD	prob.	69.36	74.81	77.14	78.62	72.63	75.41	67.16	75.65
			✓	CCCD	conf.	67.16	71.83	73.56	75.93	68.03	62.07	**74.63**	67.87
			✓	GCD	prob.	60.72	71.74	64.06	70.78	72.99	72.01	68.06	73.90
			✓	GCD	conf.	58.57	65.34	55.20	65.44	69.89	70.17	62.79	72.87
		APERTURE	✓	HSR-CCCD	prob.	**72.37**	**75.65**	**79.62**	**80.19**	73.09	**77.70**	68.65	**76.55**
		CLF-CBM	✗	CCCD	prob.	58.46	66.34	68.95	68.15	71.99	70.03	69.87	72.90
			✗	CCCD	conf.	70.74	68.62	73.71	74.36	61.65	74.65	61.80	69.82
			✗	GCD	prob.	60.75	68.06	54.53	64.82	68.61	71.87	64.54	71.01
			✗	GCD	conf.	69.88	68.21	60.78	66.53	62.40	76.23	64.28	70.10

Descriptors. CND: the class name alone. CCCD: class- and country-conditioned concept descriptors, written by Gemini 3.1 Pro and scored on the human-reviewed concept map. RAG-CCCD: CCCD grounded in country-specific documents. GCD: 44 general concept descriptors shared by all classes and countries. HSR: home-scale routing, each concept read at its own scale. HCD: hierarchical concept descriptors (high: facility level, low: component level). LC: language-generated concepts. AD: attribute descriptions. *Scores.* prob.: probability of *yes* from the yes/no token logits. conf.: confidence stated by the model. cos.: cosine similarity. OT: locally aligned optimal transport. *Settings.* Class name: zero-shot with the class name. Our training-free settings use the home-scale and global image, or their blend (APERTURE). LF-CBM trains a concept bottleneck head on two countries and tests on the third.

How well does our method perform? We show the performance of our method and its variants in table 1. Aperture reaches 80.19% macro-F1 on SiFC and 76.55% on the fMoW facility subset (Ta-

ble 1). The performance is significantly better than any of the baselines. On SiFC, APERTURE beats the best-performing baseline LaZSL by more than 10 pp, on fMoW VLG-CBM, by more than 3 pp. From both datasets, we observe that combining local crops with the global view performs better than using either view alone. For example, on SiFC All, home-scale and global-scale have an F1-score of 55.09% and 78.62%, whereas combining the scale results in the best F1-score of 80.19%.

In most cases using decoding probabilities (*prob.*) is better than MLLM-reported confidence (*conf.*). We also test the performance of different ways of creating descriptors. Class- and country-conditioned descriptors (CCCD) improve the score by more than 8 pp, over global descriptors (GCD). We also further tried a RAG-based descriptor conditioning, where we use facility documentation from specific countries instead of LLM's prior knowledge, but this yields little improvement.

Can our method generalize to other regions? Across countries, APERTURE exceeds the best CLF-CBM in four of six comparisons (Table 1). It performs better in India, the USA, and China on SiFC and France on fMoW-FC, while CLF-CBM is slightly stronger in the fMoW-FC USA and Russia. These results suggest that training-free CBMs generalize better to new regions, whereas supervised CLF-CBMs may overfit to their training countries.

Table 2: CND vs. Aperture, macro-F1 (%); Δ = Aperture $-$ CND (pp).

Dataset	Country	Gemma 4			GLM			Qwen3.5			Gemini 3.1		
		CND	APERTURE	Δ	CND	APERTURE	Δ	CND	APERTURE	Δ	CND	APERTURE	Δ
SiFC	India	60.62	72.37	+11.75	39.30	53.91	+14.61	37.43	65.04	+27.61	85.20	91.53	+6.33
	USA	68.02	75.54	+7.52	48.70	69.19	+20.49	51.40	68.86	+17.46	94.35	97.88	+3.53
	China	62.89	79.62	+16.73	48.71	60.41	+11.70	44.07	57.48	+13.41	87.20	91.12	+3.92
	All	70.40	80.19	+9.79	50.12	66.55	+16.43	48.25	66.58	+18.33	90.09	95.69	+5.60
fMoW	USA	67.52	72.32	+4.80	66.11	68.07	+1.96	62.47	67.72	+5.25	80.27	76.97	−3.30
	France	74.98	77.01	+2.03	59.94	68.75	+8.81	61.31	69.08	+7.77	68.90	85.28	+16.38
	Russia	61.46	68.65	+7.19	56.96	62.73	+5.77	67.82	62.53	−5.29	83.89	91.36	+7.47
	All	73.18	76.55	+3.37	68.15	71.52	+3.37	66.82	72.22	+5.40	78.65	81.16	+2.51

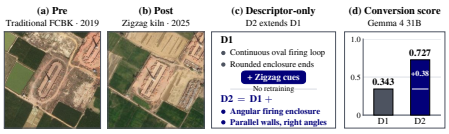


Figure 4: Descriptor-only adaptation to kiln conversion. The kiln changes from FCBK to Zigzag. Across 100 conversion sites, Zigzag-specific cues raise the conversion score from 0.343 (D1) to 0.727 (D2), with Gemma 4.

Does our method work across backbones? Combining global and home-scale views improves performance across different MLLMs, not just Gemma-4-31B (Tables 2 and 7). Each model's blended result exceeds its own best global-only concept setting wherever both were tested. This holds for Gemma 4 31B, GLM-4.6V, Qwen3.5 122B-A10B, and Gemini 3.1 Pro. The complete results with these models are in the Appendix Table 7.

Although Gemini is the strongest backbone, evaluation with proprietary models is costly and less reproducible. We therefore report our main results using open-source MLLMs. Nevertheless, results show that our training-free framework also improves over zero-shot proprietary MLLMs.

Does our method respond to visual interventions? APERTURE not only has a good performance, but it also allows better interpretability and intervention. We study whether our method responds to visual interventions that remove class-defining parts. To test this, we select 90 images across six classes, we inpaint three parts one by one, and we measure the true-class score drop from the intact image. After the first removal, the mean drop is largest at the home scale (0.10), compared with global (0.04) and zero-shot scoring (0.03). After three removals, concept-based scores drop more than zero-shot scores on average (0.43–0.49 versus 0.36; Table 3). These results show that APERTURE is *more sensitive to removed visual evidence*, with home-scale scoring responding strongest to part removal. We also visualize individual images in the appendix figs. 5 to 10.

Can descriptors adapt to technical change? Along with the generated visual interventions, we also test a real-world use case. We test whether descriptor updates can be used to track technological updates to facilities without retraining. Efforts to reduce brick-kiln pollution motivated conversion from FCBK-type to cleaner Zigzag technology (Patel et al., 2025). Across 100 conversion sites (Mondal et al., 2026), we compare original descriptors (D1) with descriptors extended by Zigzag-specific cues (D2). The conversion score increases from 0.343 to 0.727 (Fig. 4), suggesting that textual interventions can improve recognition of such changes.

Does our method work on the non-satellite CBM benchmark? We also evaluate APERTURE on CUB (Wah et al., 2011), a standard internet-image CBM benchmark. On a 600-image, 20-class subset, our method outperforms standard CbD by 5 pp but remains below CLIP-based LaZSL (68.68%; Table 4). We attribute this gap to LaZSL's random-crop strategy, which is better suited to datasets without a meaningful home scale.

Table 3: **Cumulative part-removal interventions.** Mean true-class score drop from the intact image over 15 images per class. S1–S3 cumulatively remove one, two, and three parts, respectively (listed below). The framed block averages six classes; bars end at S3, with cuts marking S1 and S2.**Bold**: largest drop per column; blue: our settings.

Method	Coal power			Sewage			Oil refinery			LNG terminal			Nuclear			Iron & steel			Mean		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Zero-shot	0.07	0.07	0.26	−0.07	0.07	0.78	0.00	0.12	0.33	**0.19**	0.14	0.59	0.04	0.23	0.28	−0.03	0.00	−0.07	0.03	0.10	0.36
Global view	0.05	0.22	0.26	0.04	0.14	**0.84**	0.01	0.22	**0.70**	0.02	0.20	**0.73**	**0.09**	**0.25**	0.35	0.02	0.04	0.07	0.04	0.18	0.49
Home scale	**0.14**	**0.36**	**0.45**	**0.19**	**0.35**	0.72	**0.01**	**0.25**	0.54	0.13	**0.31**	0.54	0.09	0.21	**0.47**	**0.06**	**0.15**	**0.23**	**0.10**	**0.27**	**0.49**
APERTURE	0.05	0.21	0.25	0.06	0.16	0.75	0.01	0.18	0.58	0.03	0.19	0.62	0.07	0.22	0.33	0.02	0.04	0.06	0.04	0.17	0.43

Parts erased. Coal power: S1 = chimney; S2 = S1 + cooling tower; S3 = S2 + coal stockpile. **Sewage:** S1 = sludge digester; S2 = S1 + aeration basin; S3 = S2 + circular clarifier. **Oil refinery:** S1 = jetty; S2 = S1 + fixed-roof tanks; S3 = S2 + floating-roof tanks. **LNG terminal:** S1 = regasification area; S2 = S1 + jetty; S3 = S2 + LNG tanks. **Nuclear:** S1 = turbine hall; S2 = S1 + cooling tower; S3 = S2 + reactor dome. **Iron & steel:** S1 = blast furnace; S2 = S1 + raw material stockyard; S3 = S2 + long mill building.

Table 4: Training-free CUB-20: macro-F1 (%) on 600 test images.

Model	Setting	F1
CLIP L/14	Zero-shot	57.69
	CBD	59.80
	LaZSL	68.68
Gemma 4 31B	Class name	34.62
	4-crop CCD	35.17
	16-crop CCD	35.17
	Global CCD	62.46
	Global + 4 crops	65.81
	Global + 16 crops	64.24

Table 5: Crop scale on SiFC. Macro-F1 (%); **All** is pooled.

Crop/view	India	USA	China	All
Fixed 2048	45.3	62.5	50.3	**54.9**
Fixed 1024	39.2	59.9	49.2	**51.1**
Fixed 512	24.7	52.6	40.1	**40.6**
Assigned	46.7	60.7	42.1	**51.3**
Routed	41.7	66.1	45.6	**55.1**
Global	69.4	74.8	77.1	**78.6**

Table 6: Training setups on SiFC. Macro-F1 (%); **All** is pooled.

Setting	Desc.	India	USA	China	All
In-domain	CCCD	74.99	88.04	82.56	**84.53**
	GCD	71.14	87.04	82.95	**83.98**
LF-CBM	CCCD	58.46	66.34	68.95	**68.15**
	GCD	60.75	68.06	54.53	**64.82**
APERTURE	HSR-CCCD	72.37	75.65	79.62	**80.19**

5.3 ABLATION

What happens if we fix the home scale? We use Gemma 4 31B on 800 SiFC sites. Fixed crops receive 113 questions each; assigned and routed views use scale-matched concepts. Assigned maps 256-px concepts to 512-px crops and 4096-px concepts to the global view; the full 4096-px tile is resized to 224 px. Fixed 2048 performs best (54.9% macro-F1), close to routed home scale (55.1%); assigned reaches 51.3% and smaller fixed crops perform worse. Global reaches 78.6%, so context remains essential and routing adds little beyond the best fixed crop (Table 5).

How well does the model perform in-domain? In-domain heads use frozen concept probabilities and five-fold within-country out-of-fold evaluation, with inner 3-fold cross-validation selecting C. LF-CBM trains on two countries and tests on the third; APERTURE is training-free. CCCD and GCD reach 84.53% and 83.98% macro-F1, versus 80.19% for training-free APERTURE. These in-domain gains require labels from the evaluated country; LF-CBM instead tests cross-country (Table 6). Remaining ablations related to alternative search strategies, etc., are present in the Appendix.

6 CONCLUSIONS

We present APERTURE , a training-free concept bottleneck model for interpretable recognition in satellite images. We show that both replacing contrastive VLMs with MLLMs and performing recognition at both global and characteristic home-scale results in significant improvement of interpretable training-free models. We also introduce SiFC, a concept-centric recognition, and show that APERTURE has much better out-of-region generalization than even supervised concept bottlenecks. We also show that our method enables easy intervention and concept-level updates. More broadly, our results suggest that effective concept-based recognition in expert domains requires reasoning about *what* visual evidence defines a concept and *where* to look for it.

Limitations and future work A limitation is that strong performance still requires blending home-scale predictions with global scores. Ideally, for component-level interpretability, home-scale evidence alone should suffice. Future work will explore strategies reducing reliance on global context.

AI USE STATEMENT

In this work, we used AI tools to assist with grammar and spelling corrections, as well as sentence refinement. We have reviewed all the AI-assisted text, and we take responsibility for the final content of this work, including the text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work is intended to support academic research and environmental monitoring by NGOs Esri (n.d.). Our dataset release will contain no satellite images, tiles, or crops. We will provide facility coordinates sourced from public government records, source references, and a Python script for authorized imagery retrieval. Prior environmental studies have used Esri imagery for land-cover annotation and permafrost-change validation (Lesiv et al., 2025; Nitze et al., 2025). Users must obtain imagery directly from Esri and comply with applicable ArcGIS and third-party terms, including download, use, and attribution requirements (Esri, 2025; Waterman, 2023). Research or nonprofit status does not remove these obligations, and our code grants no rights to the imagery

REPRODUCIBILITY STATEMENT

An anonymous downloadable code can be accessed through the link presented in the abstract. Additionally, information about the implementation and experimental setup is described in section 5.1. The details of processing the dataset will be released as easy-to-use scripts upon acceptance.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 10161–10170. IEEE, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 2018. doi: 10.1109/WACV.2018.00097.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *NeurIPS*, 32, 2019.

Shiming Chen, Bowen Duan, Salman Khan, and Fahad Shahbaz Khan. Interpretable zero-shot learning with locally-aligned vision-language model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 478–487. IEEE, 2025.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 6172–6180. IEEE, 2018.

Yang Chu, Minchao Ye, and Yuntao Qian. Fine-grained image recognition methods and their applications in remote sensing images: A review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:19640–19667, 2024.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in neural information processing systems*, 35:21400–21413, 2022.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. *Advances in Neural Information Processing Systems*, 36:37849–37875, 2023.

Esri. Gis software for mapping and spatial analytics. `https://www.esri.com/en-us/home`. Accessed September 26, 2026.

Esri. Master agreement: Products and services. `https://assets.esri.com/content/dam/esrisites/en-us/media/legal/ma-full/ma-full.pdf`, 2025. August 1, 2025. Accessed September 26, 2026.

Esri. Esri Nonprofit Program. `https://www.esri.com/en-us/industries/nonprofit/nonprofit-program`, n.d. Accessed September 26, 2026.

Patrizio Giovannotti and Alexander Gammerman. Calibrated large language models for binary question answering. *arXiv preprint arXiv:2407.01122*, 2024.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Oscar Hill, Mateo Espinosa Zarlenga, and Mateja Jamnik. Hierarchical concept-based interpretable models. In *International Conference on Learning Representations*, volume 2026, pp. 11581–11609, 2026.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025.

Yuansheng Hua, Lichao Mou, Jianzhe Lin, Konrad Heidler, and Xiao Xiang Zhu. Aerial scene understanding in the wild: Multi-scene recognition via prototype-based memory networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:89–102, 2021.

Yuansheng Hua, Jiasong Zhu, and Qingquan Li. Pcinet: a prototype-and concept-based interpretable network for mutli-scene recognition. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:265–270, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Ioannis Kakogeorgiou and Konstantinos Karantzalos. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103:102520, 2021.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes! *ICCV*, 2025.

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16521–16540. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kim23g.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Myroslava Lesiv, Steffen Fritz, Martina Dürauer, Ivelina Georgieva, Marcel Buchhorn, Luc Bertels, Nandika Tsendbazar, Ruben Van De Kerchove, Daniele Zanaga, Dmitry Schepaschenko, et al. A global reference data set for land cover mapping at 10 m resolution. *Earth System Science Data Discussions*, 2025:1–9, 2025.

Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International journal of applied earth observation and geoinformation*, 124:103497, 2023.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 6535–6563, 2026.

Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Field-guide-inspired zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Utkarsh Kumar Mall, Cheng Perng Phoo, Meilin Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. In *International Conference on Learning Representations*, volume 2024, pp. 49294–49314, 2024.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jlAjNL8z5cs.

Rishabh Mondal, Jeet Parab, Heer Kubadia, Shataxi Dubey, Shardul Junagade, Zeel Bharatkumar Patel, and Nipun Batra. Sentinelkilndb: A large-scale dataset and benchmark for obb brick kiln detection in south asia using satellite imagery. *Advances in Neural Information Processing Systems*, 38, 2026.

Ingmar Nitze, Konrad Heidler, Nina Nesterova, Jonas Küpper, Emma Schütt, Tobias Hölzer, Sophia Barth, Mark J Lara, Anna K Liljedahl, and Guido Grosse. Darts: Multi-year database of ai-detected retrogressive thaw slumps in the circum-arctic permafrost region. *Scientific Data*, 12(1): 1512, 2025.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Konstantinos P Panousis, Dino Ienco, and Diego Marcos. Coarse-to-fine concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:105171–105199, 2024.

Zeel B Patel, Rishabh Mondal, Shataxi Dubey, Suraj Jaiswal, Sarath Guttikunda, and Nipun Batra. Space to policy: Scalable brick kiln detection and automatic compliance monitoring with geospatial data. *ACM Journal on Computing and Sustainable Societies*, 3(3):1–29, 2025.

Federico Pittino, Vesna Dimitrievska, and Rudolf Heer. Hierarchical concept bottleneck models for vision and their application to explainable fine classification and tracking. *Engineering Applications of Artificial Intelligence*, 118:105674, 2023.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023.

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

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4065–4076, 2023. doi: 10.1109/ICCV51070.2023.00378.

Shreelekha Revankar, Cheng Perng Phoo, Utkarsh Kumar Mall, Bharath Hariharan, and Kavita Bala. Scale-aware recognition in satellite images under resource constraints. In *International Conference on Learning Representations*, volume 2025, pp. 54725–54737, 2025.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37:79057–79094, 2024.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pp. 5901–5904. IEEE, 2019.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336. IEEE, 2025.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *AAAI*, 2024.

Robert Waterman. World Imagery Wayback Export. `https://www.esri.com/arcgis-blog/products/arcgis-living-atlas/imagery/wayback-export`, 2023. Accessed September 26, 2026.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2018. doi: 10.1109/CVPR.2018.00418.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Shin'ya Yamaguchi, Kosuke Nishida, Daiki Chijiwa, and Yasutoshi Ida. Zero-shot concept bottleneck models. *arXiv preprint arXiv:2502.09018*, 2025.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19187–19197. IEEE, 2023.

Liang Yao, Fan Liu, Hongbo Lu, Chuanyi Zhang, Rui Min, Shengxiang Xu, Shimin Di, and Pai Peng. Remotereasoner: Towards unifying geospatial reasoning workflow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11883–11891, 2026.

A Appendix

A.1 Changes to Concept Questioning

From Attribute Phrases to Structured Descriptors.

We observe that instruction-following MLLMs respond more reliably to complete sentences with explicit visual cues 1. We therefore expand each descriptor into a presence with *appearance, spatial placement, and likely confusers.* e.g., instead of using only "`cooling tower`", we ask, "`Is a large hyperbolic concrete shell with a thick rim and dark open center visible near the turbine hall? Do not confuse it with a storage dome, which has a solid roof and casts a shorter shadow.`". This structure provides both supporting and contrasting evidence, making concept verification less ambiguous. The goal is not merely utilizing larger context windows but providing a clearer description of what the model should identify and exclude.

Each descriptor combines visual appearance with cues that distinguish similar objects. The generation template below is reconstructed from the saved descriptor structure; the example retains the saved question wording with its scale placeholder filled.

A. Descriptor-generation template (reconstructed)

Input concept: {`concept`}

Expand this concept into a structured description for recognition in overhead satellite imagery. Focus on the object itself without naming the facility class.

From Explicit Class Names to Class-Conditioned Descriptors The original CBD implementation prepends class names to the descriptors. Prepending class names provides better context to a descriptor. For example, a "leg of a table" semantically is a very different concepts from "leg of a lion". This results in improved performance but affects interpretability, as the final classification is not just a function of the concept bottleneck, and some indirect signal is coming from the class name encoded. Therefore, rather than prepending the class name, we condition the concept descriptors on the class name. Building on the previously derived structured descriptions, we condition each concept on the facility class, because the same component can appear differently across classes. Prepending labels not only limits interpretability, but in our experiments, prepending labels also reduces performance. We construct $d_{k|y}$, a visual description of concept k conditioned on class y, while omitting the class name from the verification question. A cooling-tower descriptor can be ``a cooling tower beside boiler buildings and coal stockpiles'' for power plants or ``a cooling tower beside a domed building" for nuclear power plants. This captures class-specific appearance while keeping verification focused on visible evidence.

Appearance: Describe its visible shape, approximate size, colour, surface features, and shadow.

Confusers: Identify similar-looking objects and the visible cues that distinguish them.

Verification question: Assemble the fields as follows:

This is an aerial satellite image, {`scale_text`}. Is a/an {`concept`} visible in it? {`appearance`} Do not confuse with: {`confusers`} Answer with exactly one word: yes or no.

B. Worked example: Chimney stack

Input concept: Chimney stack. **Scale text:** about 2 km across.

Verification question: This is an aerial satellite image, about 2 km across. Is a chimney stack visible in it? Appearing as a small 2 to 10 meter circular dot in gray, white, or red-and-white stripes, a chimney stack is most easily identified by its exceptionally long, needle-like shadow cast across the ground. Do not confuse with: It is most often mistaken for silos, water towers, or flare stacks, but can be distinguished by its much longer, narrower shadow and the occasional presence of a visible smoke or steam plume emitting from the open top. Answer with exactly one word: yes or no.

From Explicit Class Names to Class-Conditioned Descriptors. The class guides the description of appearance, confusers, and placement. The adapted verification question below uses these visual cues without explicitly naming the class.

C. Class-conditioned generation template (condensed)

Write visual descriptors for overhead satellite imagery at approximately 0.5 m per pixel.

Concept: {concept}

Facility class: {class}

Generic description: Appearance: {appearance}; confusers: {confusers}; placement: {placement}.

Rewrite these fields for how the concept appears in this facility class. Preserve valid details and specialize its size, count, form, and surrounding structures where appropriate.

Return only a JSON object with three fields:

appearance: Visible shape, size, colour, shadow, and typical count.

confusers: Similar-looking objects and the visible details that distinguish them.

placement: Nearby structures and visible connections.

Constraints: Describe the class generally, not one site. Retain generic wording when no class-specific difference exists. Do not include the facility class name in the returned fields.

D. Worked example: Chimney stack (adapted question)

Generation inputs: Concept: chimney stack; class: coal or thermal power plant.

Class-name-free verification question: This is an aerial satellite image, about 2 km across. Is a chimney stack visible in it? Appearing as one to four bright circular dots 15 to 30 m across at the base, these structures are most easily identified by their very long, needle-thin shadows of 100 to 300 m lying across the site and the frequent presence of a visible plume. Do not confuse with: Cooling towers also emit plumes but are much wider with shorter, thicker shadows, while transmission towers cast shorter, lattice-patterned shadows and stand in linear networks. Placement: Positioned immediately beside the main boiler building or emissions control equipment, it is connected to them by large, visible flue gas ducts. Answer with exactly one word: yes or no.

F. Worked example: The same concept across three countries

Shared generation inputs: Concept: chimney stack; class: coal or thermal power plant. Only the country input changes.

Shared verification format: This is an aerial satellite image, about 260 m across. Is a chimney stack visible in it? {appearance} Do not confuse with: {confusers} Answer with exactly one word: yes or no.

Country-conditioned appearance excerpts:

India. Appearing as **one to six** bright, **concrete-grey** circular dots 15 to 30 m across at the base, these structures are identifiable by their extremely long, needle-thin shadows of **150 to 300 m** and the frequent presence of a visible plume.

China. Appearing as **one to four** bright **white or grey** circular dots 15 to 30 m across, these structures are most easily identified by their extremely long, needle-thin shadows of **150 to over 300 m** lying across the site and the frequent presence of a **dense white plume**.

USA. Appearing as **one to three** bright, **light grey** circular dots 15 to 30 m across at the base, these structures are most easily identified by their very long, needle-thin shadows of **150 to over 300 m** lying across the site and the frequent presence of a visible plume **if the unit remains active**.

Reading guide: Country labels identify generation inputs and are not included in the object-verification question. Bold highlights differences in the saved LLM-generated descriptions, not measured country-level statistics. Only the appearance field is shown for comparison.

Concept-Specific Home-Scale Assignment. The home scale specifies the source-image crop width used to verify a concept. The template below is reconstructed; the example reports a saved assignment.

G. Home-scale assignment template (reconstructed)

Concept: {concept}

Descriptor: Appearance: {appearance}; confusers: {confusers}; placement: {placement}.

Image setup: The source image is 4096×4096 pixels at approximately 0.5 m per pixel. Each selected crop is resized to 224×224 pixels before verification.

Task: Assign a crop width that balances visible detail with the surrounding context needed to recognize this concept.

Choose from 4096, 2048, 1024, 512, or 256 source pixels. Use wider crops for large structures or spatial arrangements, and tighter crops for small components. Consider the visual cues needed to distinguish the concept from similar objects.

Return only: A JSON object with the field home_scale_px containing one of the allowed widths.

H. Worked example: Chimney stack

Descriptor entry: Concept: chimney stack; class: coal or thermal power plant; country: India.

Descriptor cues (condensed): Small, concrete-grey circular bases with long, needle-thin shadows, positioned beside boiler buildings and connected by visible flue gas ducts.

Saved assignment: {"home_scale_px": 512}

Use during inference: Greedy quadtree routing stops at a 512×512-pixel crop from the original image. This crop is resized to 224×224 pixels for concept verification.

Note: The home scale is an assigned setting, not an experimentally established optimum.

Final Inference Prompts for APERTURE. Each descriptor provides questions for global verification, quadrant routing, and home-scale verification. The templates below preserve the saved prompt structure; placeholders are filled from the country-selected descriptor entry.

I. Global verification

This is an aerial satellite image, about 2 km across. Is a/an {concept} of the kind found at a/an {class_name} visible in it? {appearance} Do not confuse with: {confusers} Placement: {placement} Answer with exactly one word: yes or no.

J. Quadrant routing

This aerial satellite image covers {scale_text} and is divided into four equal quadrants: 1 upper-left, 2 upper-right, 3 lower-left, and 4 lower-right. Assume the target is somewhere in this crop. Which ONE quadrant contains the clearest visual evidence consistent with the concept {concept} associated with {class_name}?

Target appearance: {appearance}

Expected placement: {placement}

Do not confuse with: {confusers}

This is only a zoom-location decision, not a final yes-or-no presence decision. If the evidence is weak, still choose the quadrant containing the strongest candidate evidence. Answer with exactly one digit: 1, 2, 3, or 4.

K. Home-scale verification: object-level concepts

This is an aerial satellite image, {scale_text}. Is a/an {concept} visible in it? {appearance} Do not confuse with: {confusers} Answer with exactly one word: yes or no.

Routing uses a numbered grid; home-scale verification uses the clean crop. The scale text follows the current crop width: 4096, 2048, 1024, 512, and 256 pixels correspond to "about 2 km across", "about 1 km across", "about 500 m across", "about 260 m across", and "about 130 m across",

respectively. For tile-level concepts, the saved home question retains the global question structure with the appropriate scale text. A home scale of 4096 reuses the global evidence.

L. Saved example: Chimney stack in India

Descriptor metadata: Concept: chimney stack; class: coal or thermal power plant; country: India; home scale: 512 pixels.

Routing scales: $4096 \rightarrow 2048 \rightarrow 1024 \rightarrow 512$. The routing question is asked at the first three scales, followed by verification at 512 pixels.

Saved home-scale question:

This is an aerial satellite image, about 260 m across. Is a chimney stack visible in it? Appearing as one to six bright, concrete-grey circular dots 15 to 30 m across at the base, these structures are identifiable by their extremely long, needle-thin shadows of 150 to 300 m and the frequent presence of a visible plume. Do not confuse with: Cooling towers also emit plumes but are much wider with massive dark central openings and shorter, thicker shadows, while transmission towers cast lattice-patterned shadows and stand in linear networks. Answer with exactly one word: yes or no.

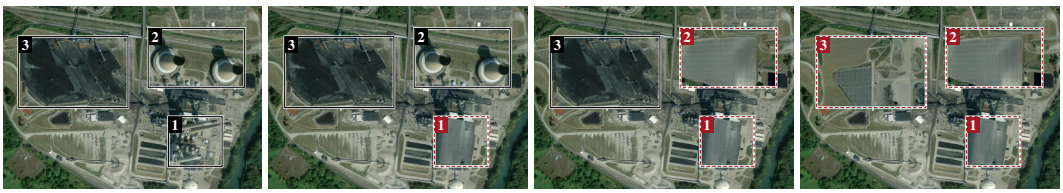
What if we use a different search strategy?

We test wider search using frozen Gemma 4 31B, keeping the descriptors, home scales, and SiFC images unchanged. We expand greedy top-1 quadtree routing to beams of 2, 3, and 4 (Table 11). Home denotes home-only scoring; Blend combines global and home views. Blend settings were selected using evaluation labels. Home-only macro-F1 rises from 55.09% to 64.13% with beam 3. The global–home blend changes little: beam 2 reaches 80.42%, only 0.23 percentage points above greedy (80.19%), while beams 3 and 4 score lower. Wider beams add crop evaluations with little blend gain, so we retain greedy top-1 routing.

Gemma 4 31B *Cumulative component removal*

Zero-shot uses class-answer probability; concept methods use originally present concepts ($p_{\text{yes}} > 0.5$, $\tau = 30$).

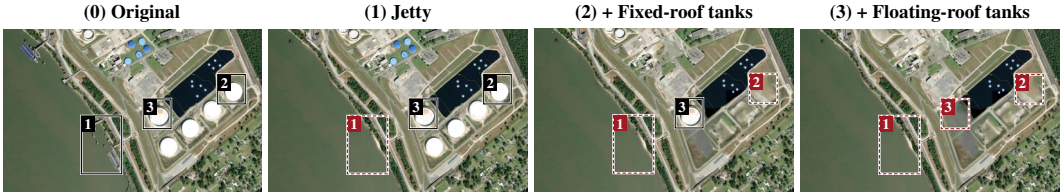
Coal power plant



Method	S_0	S_1	Δ_1	S_2	Δ_2	S_3	Δ_3
Zero-shot	0.9999	0.9333	0.0666	0.9333	0.0666	0.7396	0.2603
Global	0.9811	0.9293	0.0518	0.7656	0.2156	0.7227	0.2584
Home	0.9736	0.8291	**0.1446**	0.6108	**0.3629**	0.5209	**0.4528**
APERTURE	0.9736	0.8291	**0.1446**	0.6108	**0.3629**	0.5209	**0.4528**

Figure 5: One selected site; S_j averages 15 images after step j. Boxes mark representative removals. Red $\Delta_j = S_0 - S_j$; bold: largest concept-method drop.

Oil refinery



Method	S_0	S_1	Δ_1	S_2	Δ_2	S_3	Δ_3
Zero-shot	0.9999	0.9993	0.0007	0.8830	0.1169	0.6689	0.3310
Global	0.9531	0.9397	0.0134	0.7334	0.2197	0.2538	**0.6993**
Home	0.9778	0.9640	**0.0138**	0.7291	**0.2487**	0.4375	0.5403
APERTURE	0.9531	0.9397	0.0134	0.7334	0.2197	0.2538	**0.6993**

Figure 6: One selected site; S_j averages 15 images after step j. Boxes mark representative removals. Red $\Delta_j = S_0 - S_j$; bold: largest concept-method drop.

Nuclear power plant

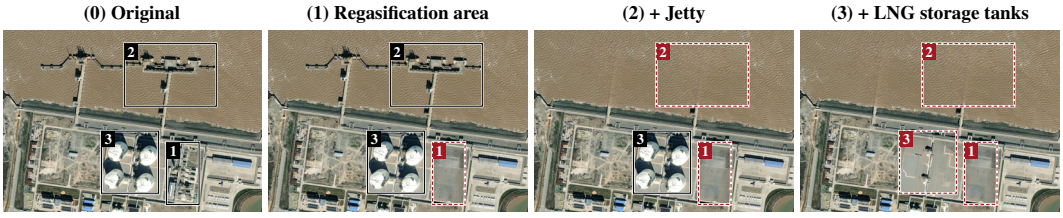


Method	S_0	S_1	Δ_1	S_2	Δ_2	S_3	Δ_3
Zero-shot	0.6493	0.6107	0.0386	0.4233	0.2260	0.3712	0.2782
Global	0.9290	0.8417	**0.0872**	0.6765	**0.2524**	0.5801	0.3489
Home	0.9486	0.8627	0.0859	0.7435	0.2051	0.4741	**0.4746**
APERTURE	0.9486	0.8627	0.0859	0.7435	0.2051	0.4741	**0.4746**

Figure 7: One selected site; S_j averages 15 images after step j. Boxes mark representative removals. Red $\Delta_j = S_0 - S_j$; bold: largest concept-method drop.

Gemma 4 31B *Cumulative component removal*

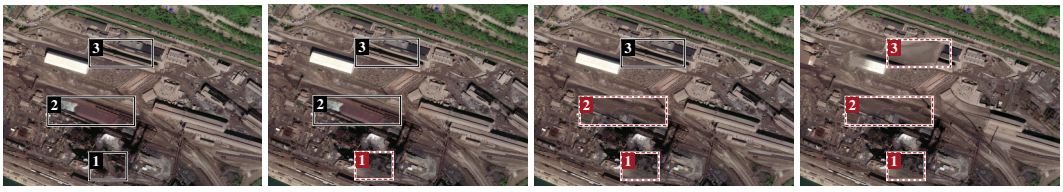
LNG terminal



Method	S_0	S_1	Δ_1	S_2	Δ_2	S_3	Δ_3
Zero-shot	0.9329	0.7388	0.1941	0.7944	0.1384	0.3464	0.5864
Global	0.9629	0.9457	0.0172	0.7582	0.2048	0.2317	0.7312
Home	0.9684	0.8405	**0.1279**	0.6534	**0.3149**	0.4292	0.5392
APERTURE	0.9615	0.9447	0.0169	0.7558	0.2057	0.2300	**0.7315**

Figure 8: One selected site; S_j averages 15 images after step j. Boxes mark representative removals. Red $\Delta_j = S_0 - S_j$; bold: largest concept-method drop.

Iron and steel plant



Method	S_0	S_1	Δ_1	S_2	Δ_2	S_3	Δ_3
Zero-shot	0.7997	0.8343	-0.0345	0.8002	-0.0005	0.8665	-0.0668
Global	0.9674	0.9482	0.0192	0.9269	0.0405	0.8957	0.0717
Home	0.9681	0.9076	**0.0605**	0.8140	**0.1541**	0.7372	**0.2309**
APERTURE	0.9681	0.9076	**0.0605**	0.8140	**0.1541**	0.7372	**0.2309**

Figure 9: One selected site; S_j averages 15 images after step j. Boxes mark representative removals. Red $\Delta_j = S_0 - S_j$; bold: largest concept-method drop.

Sewage treatment plant



Method	S_0	S_1	Δ_1	S_2	Δ_2	S_3	Δ_3
Zero-shot	0.9333	0.9999	-0.0665	0.8665	0.0668	0.1528	0.7806
Global	0.9567	0.9131	0.0436	0.8176	0.1392	0.1199	0.8368
Home	0.9742	0.7823	**0.1919**	0.6273	**0.3470**	0.2520	0.7223
APERTURE	0.9841	0.9273	0.0569	0.8096	0.1745	0.1188	**0.8654**

Figure 10: One selected site; S_j averages 15 images after step j. Boxes mark representative removals. Red $\Delta_j = S_0 - S_j$; bold: largest concept-method drop.

Table 7: Macro F1 (%) on SiFC (six classes, 800 images; India, USA, China) and on the fMoW subset (USA, France, Russia). All rows share one MLLM and are training-free (TF ✓) except the CBM rows, which fit a linear head with leave-one-country-out folds (✗). Within each model block, **bold** is the best and underline the second best per column. The row in blue is our full method.

					SiFC (ours)				fMoW			
Model	**Setting**	**Descriptor**	**Score**	**TF**	**India**	**USA**	**China**	**All**	**USA**	**France**	**Russia**	**All**
Baseline	CLIP zero-shot	CND	cos.	✓	53.03	**67.19**	53.29	62.16	**71.29**	57.71	51.91	65.85
	CBD	CBD	cos.	✓	59.37	66.60	57.73	66.42	70.57	54.11	51.02	62.74
	LaZSL	AD	OT	✓	69.38	**72.83**	58.78	**69.76**	**78.63**	59.86	57.17	69.90
	CF-CBM (high)	HCD	cos.	✗	40.25	25.67	48.60	39.35	12.60	25.89	24.11	23.18
	CF-CBM (low)	HCD	cos.	✗	53.92	40.87	60.69	49.88	41.30	**72.52**	58.17	55.92
	LaBo	LC	cos.	✗	66.70	62.97	53.39	60.44	64.75	57.50	**65.21**	65.11
	VLG-CBM	LC	logit	✗	**71.91**	64.04	**66.88**	**68.21**	69.40	68.26	62.28	**71.62**
Gemma 4 31B	Class name only	CND	–	✓	60.62	68.02	62.89	70.40	67.52	74.98	61.46	73.18
	Home scale only	HSR-CCCD	prob.	✓	41.74	66.10	45.56	55.09	57.76	**77.70**	49.04	63.46
	Global concepts	RAG-CCCD	prob.	✓	70.68	**75.65**	74.67	77.53	69.69	75.17	70.07	74.40
		CCCD	prob.	✓	69.36	74.81	77.14	78.62	72.63	75.41	67.16	75.54
		CCCD	conf.	✓	67.16	71.83	73.56	75.93	68.03	62.07	**74.63**	67.87
		GCD	prob.	✓	60.72	71.74	64.06	70.78	**73.05**	72.01	68.06	73.90
		GCD	conf.	✓	58.57	65.34	55.20	65.44	69.89	70.17	62.79	72.87
	Global + home (ours)	HSR-CCCD	prob.	✓	**72.37**	75.54	**79.62**	**80.19**	72.32	77.01	68.65	**76.55**
	LF-CBM	CCCD	prob.	✗	58.46	66.34	68.95	68.15	72.99	70.03	69.87	72.90
		CCCD	conf.	✗	70.74	68.62	73.71	74.36	61.65	74.65	61.80	69.82
		GCD	prob.	✗	60.75	68.06	54.53	64.82	68.61	71.87	64.54	71.01
		GCD	conf.	✗	69.88	68.21	60.78	66.53	62.40	76.23	64.28	70.10
GLM-4.6V	Class name only	CND	–	✓	39.30	48.70	48.71	50.12	66.11	59.94	56.96	68.15
	Home scale only	HSR-CCCD	prob.	✓	30.62	54.22	33.20	40.97	62.59	66.27	54.12	63.97
	Global concepts	CCCD	prob.	✓	52.16	63.63	60.89	65.02	65.50	60.04	58.97	67.10
		CCCD	conf.	✓	46.83	56.47	57.45	60.62	61.26	54.55	34.42	56.84
		GCD	prob.	✓	41.32	52.80	47.66	53.61	62.48	61.42	52.76	65.42
		GCD	conf.	✓	35.59	50.08	36.73	46.29	61.22	48.70	45.79	55.61
	Global + home (ours)	HSR-CCCD	prob.	✓	53.91	**69.19**	60.41	**66.55**	**68.07**	**68.75**	**62.73**	**71.52**
	LF-CBM	CCCD	prob.	✗	52.66	57.99	64.71	62.60	63.91	52.46	52.41	62.80
		CCCD	conf.	✗	51.33	62.29	59.18	60.30	66.24	57.89	56.84	64.08
		GCD	prob.	✗	**55.88**	67.03	**65.61**	66.54	63.45	56.39	50.37	63.89
		GCD	conf.	✗	47.88	62.71	55.69	59.64	66.24	57.89	56.84	64.08
Qwen3.5 122B-A10B	Class name only	CND	–	✓	37.43	51.40	44.07	48.25	62.47	61.31	**67.82**	66.82
	Home scale only	HSR-CCCD	prob.	✓	40.65	55.11	42.48	48.52	54.52	64.06	52.95	58.99
	Global concepts	CCCD	prob.	✓	58.65	62.60	53.36	61.78	67.49	66.68	59.52	70.69
		CCCD	conf.	✓	60.05	62.55	55.26	63.24	54.52	63.21	52.95	58.99
		GCD	prob.	✓	49.07	57.25	41.05	53.11	**70.60**	61.54	61.28	69.70
		GCD	conf.	✓	38.72	44.18	40.56	45.82	65.44	49.60	43.96	56.09
	Global + home (ours)	HSR-CCCD	prob.	✓	**65.04**	**68.86**	57.48	**66.58**	67.72	69.08	62.53	**72.22**
	LF-CBM	CCCD	prob.	✗	57.12	68.61	51.11	61.77	69.16	59.07	65.30	68.97
		CCCD	conf.	✗	56.00	61.91	**63.69**	64.06	68.40	56.84	58.80	66.53
		GCD	prob.	✗	59.80	63.83	50.06	59.74	63.23	**70.55**	61.21	68.37
		GCD	conf.	✗	47.02	57.45	48.85	54.63	69.77	56.98	56.95	67.79
Gemini 3.1 Pro	Class name only	CND	–	✓	85.20	94.35	87.20	90.09	80.27	68.90	83.89	78.65
	Global scale	CCCD	conf.	✓	86.78	90.01	91.12	92.53	73.28	80.31	89.32	78.32
	Home scale only	HSR-CCCD	conf.	✓	72.23	88.89	77.85	74.66	70.09	76.09	77.26	72.39
	Global + home (ours)	HSR-CCCD	conf.	✓	91.53	97.88	91.12	95.67	**76.97**	**85.28**	**91.36**	**81.16**

Descriptors. CND: the class name n_y alone. CCCD: class- and country-conditioned concept descriptors $d_{k|y,r}$ for concept k, class y and country r, written by Gemini 3.1 Pro (v4) and scored on the human-reviewed concept map. RAG-CCCD: CCCD grounded in country-specific documents. GCD: 44 general concept descriptors d_k shared by all classes and countries. HSR: home-scale routing, where each concept is read at its own scale. **Scores.** prob.: probability of *yes* from the yes/no token logits. conf.: confidence stated by the model in its answer. **Gemini 3.1 Pro** is a closed-source reference. **Settings.** Class name only: zero-shot with n_y. Home scale only: HSR without the global view. Global concepts: the full tile at 224 px. Global + home: both views. CBM (LOCO): a concept bottleneck head trained on two countries and tested on the third. HCD: hierarchical concept descriptors, with facility-level (high) and component-level (low) descriptions. LC: language-generated concepts expressed as short visual phrases. AD: attribute descriptions. OT: locally aligned optimal-transport score.

Table 8: Proposed class and country distribution of the fMoW-FC facility subset.

Class	USA	France	Russia	Total
Shipyard	15	6	5	26
Solar farm	24	20	6	50
Storage tank	74	66	47	187
Water treatment facility	26	43	27	96
Wind farm	22	14	5	41
Total	**161**	**149**	**90**	**400**

Table 9: SiFC facility sites by class and country.

Facility class	India	USA	China	Total
Iron and steel plant	48	28	104	180
LNG terminal	8	13	28	49
Nuclear power plant	7	52	15	74
Oil refinery	24	70	65	159
Power plant	101	42	37	180
Sewage treatment plant	22	90	46	158
Total	**210**	**295**	**295**	**800**

Table 10: Government references for SiFC facility identity and locality. individual verification.

Class	India	USA	China
Iron and steel	PIB	EPA	MEE
LNG terminals	PNGRB	FERC	MEE
Nuclear plants	NPCIL	EIA-860	NNSA
Oil refineries	PPAC	EIA-820	MEE
Power plants	CEA	EIA-860	MEE
Sewage treatment	CPCB	EPA ECHO	MEE

Table 11: **Search width on SiFC.** Macro-F1 (%); **All**: pooled.

	Top-1		Beam-2		Beam-3		Beam-4	
Region	Home	Blend	Home	Blend	Home	Blend	Home	Blend
India	41.74	72.37	49.13	71.26	53.04	71.55	52.11	70.41
USA	66.10	75.54	73.13	77.91	76.40	78.27	74.00	77.94
China	45.56	79.62	53.17	80.48	53.31	79.01	53.84	79.67
All	**55.09**	**80.19**	**61.58**	**80.42**	**64.13**	**80.11**	**63.10**	**80.12**