TRACE: TOWARDS INTERPRETABLE CARBON EMISSION ESTIMATION VIA REASONING-TO-PREDICTION FROM SATELLITE IMAGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate monitoring of urban carbon emissions is essential for climate mitigation, yet satellite-based deep models remain difficult to interpret: they predict an emission value without revealing which scene properties support it. Vision-Language Models (VLMs) have the potential to bridge this semantic gap; however, applying them directly to carbon-emission estimation remains challenging because they are prone to visual hallucinations and lack mechanisms tailored to continuous regression. We propose TRACE (Textual Reasoning-Augmented Carbon Estimation), a textual reasoning-augmented framework compatible with different visual encoders in which an Analyst VLM generates image-grounded environmental analyses and a downstream Predictor maps visual and textual features to an emission estimate. In the absence of ground-truth analytical annotations, teacher-generated pseudo-analyses are used to initialize the Predictor, followed by alternating Analyst–Predictor optimization through Periodical Evolution. Experiments using data from five cities show improvements over image-only baselines across four visual encoders. TRACE achieves relative R^2 gains of 3.6%–6.8%, while blind human evaluation shows improved factual grounding and carbon relevance over unfine-tuned VLM analyses.

1 INTRODUCTION

Driven by rapid urbanization, cities account for the vast majority of anthropogenic greenhouse gas emissions (Crippa et al., 2021; Shen et al., 2022), so monitoring urban emissions matters for carbon-neutrality targets (UNFCCC, 2015; IPCC, 2018). Satellite remote sensing provides repeated, large-area observations at fine spatial scales (Lin et al., 2022; Uniyal et al., 2022; Liu et al., 2024c). Deep learning models have improved estimation accuracy (Javed et al., 2023; Saputra et al., 2025), but their output is usually just a number, as shown in Fig. 1(a). For many environmental policy decisions, that estimate is more useful when accompanied by a traceable account of the visible structures and contextual factors associated with it.

VLMs integrate visual and linguistic representations, enabling cross-modal interpretation of imagery and the generation of natural-language analyses of visual content (Yang et al., 2024; Zhou et al., 2024). However, directly applying VLMs in this setting presents two challenges (Fig. 1(b)). They may hallucinate objects or land-use types absent from the image or emphasize features unrelated to emissions (Augustin et al., 2025; Liu et al., 2024b). Moreover, because VLMs are optimized primarily for language generation, they are not well suited to precise continuous regression (Li et al., 2025). An interpretable carbon estimator must therefore ground its analysis in fine-grained visual evidence while retaining a separate mechanism for numerical prediction. Here, interpretability means that each estimate is accompanied by an image-grounded analysis that can be inspected before regression.

We propose **T**extual **R**easoning-**A**ugmented **C**arbon **E**stimation (TRACE), a framework that augments visual encoders with learned textual reasoning for carbon estimation (Fig. 1(c)). An Analyst VLM learns through reinforcement learning (RL) to express carbon-related reasoning as an environmental analysis of the satellite image and contextual metadata. Separate pretrained encoders map the image and analysis to visual and textual embeddings, and a multilayer perceptron (MLP) predicts the

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.

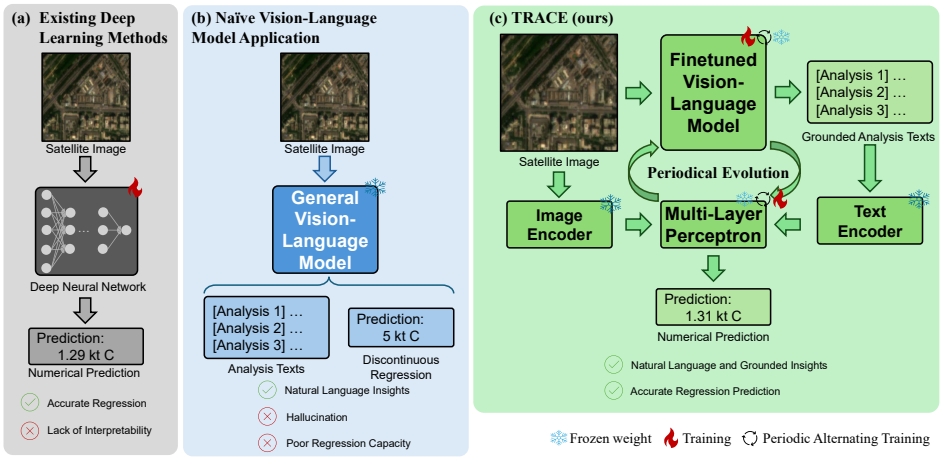


Figure 1: Comparison with TRACE. Predictions are in kt C per grid cell per month.

emission value from their concatenation. This design augments visual representations with analysis text while retaining a separate regression module.

The analysis enters the Predictor directly, so reduced regression error provides a training signal for its utility. This signal alone cannot ensure factual grounding: the Analyst may exploit a fixed Predictor's errors, while an accurate description may omit emission-relevant cues. TRACE therefore combines grounding and prediction rewards.

Training this pipeline presents two coupled challenges. First, the data contain emission values but no expert-written analyses, leaving no reference analyses for Predictor initialization. Second, a fixed Predictor may expose stable error patterns that the Analyst can exploit. Teacher-generated pseudo-analyses address the first challenge by initializing the Predictor. Periodical Evolution addresses the second by alternating Analyst optimization under grounding and prediction rewards with Predictor retraining on updated analyses. Freezing the Predictor within each Analyst stage stabilizes the reward, while updating it between stages reduces the Analyst's reliance on earlier error patterns.

Our contributions are as follows:

- We introduce a textual reasoning-augmented framework for carbon-emission estimation. TRACE combines inspectable environmental analyses with features from visual encoders and trains a separate Predictor for continuous estimation.
- We introduce Periodical Evolution to train TRACE without expert-written analysis labels. A Teacher VLM supplies pseudo-analyses to initialize the Predictor. The Analyst is then optimized with two separate rewards, while the Predictor is updated on newly generated analyses using the regression objective.
- We evaluate TRACE on satellite imagery paired with carbon-emission data from five cities. Comparisons across four visual encoders, controlled text-input ablations, and analysis-quality evaluation examine predictive gains and the utility of the generated analyses.

2 Related Work

2.1 Vision-based Carbon Emission Estimation

Carbon-emission estimates have traditionally relied on statistical proxies such as nighttime lights (Han et al., 2018; Yang et al., 2020) or on coarse-resolution transport models (Lespinas et al., 2020). High-resolution satellite products (Roy et al., 2021) now support image-based estimation with convolutional networks and vision transformers, which learn nonlinear spatial features for regression (Zhou et al., 2023; Song et al., 2024; Marjani et al., 2025). These models can be accurate, but they do not state which visual evidence supports a prediction. TRACE retains the visual regression pipeline while adding an explicit, inspectable environmental analysis.

2.2 Vision-Language Models in Remote Sensing

VLM research in remote sensing connects satellite imagery with language for retrieval and semantic localization (Zhang et al., 2024; Liu et al., 2024a), and these representations have been used to predict socioeconomic indicators (Yan et al., 2024; Hao et al., 2025). Generative models also support remote-sensing captioning and visual question answering (Bazi et al., 2024; Kuckreja et al., 2024). Urban-R1 (Wang et al., 2025) and CityRiSE (Liu et al., 2025) use reinforcement learning to improve reasoning over urban imagery, but their targets are discretized or normalized socioeconomic status. TRACE addresses monthly carbon-emission regression at 1 km resolution. It uses the generated analysis as an intermediate input to a numerical Predictor, which separates language generation from regression and makes mitigating visual hallucination a training target.

2.3 Reinforcement Learning for Visual Reasoning

Reinforcement learning from human feedback (RLHF) aligns language models with human preferences (Ouyang et al., 2022; Bai et al., 2022a), commonly through objectives related to proximal policy optimization (PPO) (Schulman et al., 2017) or direct preference optimization (DPO) (Rafailov et al., 2023). Collecting reliable preferences is expensive in scientific domains that require specialist knowledge (Chaudhari et al., 2025). Reinforcement learning with verifiable rewards instead uses outcome-based signals (Guo et al., 2025; Yue et al., 2025), while reinforcement learning from AI feedback uses model-generated judgments (Bai et al., 2022b; Xiong et al., 2025). TRACE combines automated quality judgments with prediction feedback from a carbon Predictor. During RL, the Predictor remains fixed while the Analyst is trained. The Predictor is then retrained using analyses from the updated Analyst. This alternating strategy is designed to avoid training instability by keeping the Predictor fixed during each Analyst optimization stage.

3 Methodology

3.1 Overview of TRACE

TRACE has two training phases as illustrated in Fig. 2. During cold-start initialization, a Teacher VLM generates a pseudo-analysis for each satellite image in the Predictor-training subset using its city, month, and emission label. These analyses initialize the Predictor by linking visual and text features to emission values. These analyses are not used for supervised training of the Analyst.

Periodical Evolution then trains a smaller Analyst VLM with RL on a separate Analyst-training subset. The Analysis Reward evaluates whether a response is grounded and relevant, while the Prediction Reward measures whether the response helps the frozen Predictor reduce error. After each Analyst stage, the updated Analyst regenerates analyses for the Predictor-training subset, which is reused to update the Predictor. The Predictor remains fixed during each Analyst stage, and the Analyst remains fixed during each Predictor stage.

The Teacher receives emission labels only from the Predictor-training subset. During Analyst optimization, the Analyst takes only the image and city/month metadata from the disjoint Analyst-training subset as input, while the corresponding labels are used only to compute rewards. Predictor optimization remains confined to the Predictor-training subset. At inference, the Analyst receives a satellite image and its city/month metadata; the analysis and image are then encoded for prediction.

3.2 Cold-Start Initialization

Carbon-emission datasets provide numerical labels but generally do not pair each satellite image with an expert environmental analysis. This creates a circular initialization problem: the Predictor needs text examples before it can assess whether an analysis improves regression, while the Analyst needs that assessment during RL.

To break this dependency, we use a pretrained Teacher VLM without task-specific fine-tuning to generate pseudo-analyses. These texts provide the initial language input needed to train the Predictor, but may contain errors inherited from the Teacher. They are used only to initialize the Predictor, which subsequently provides the initial Prediction Reward.

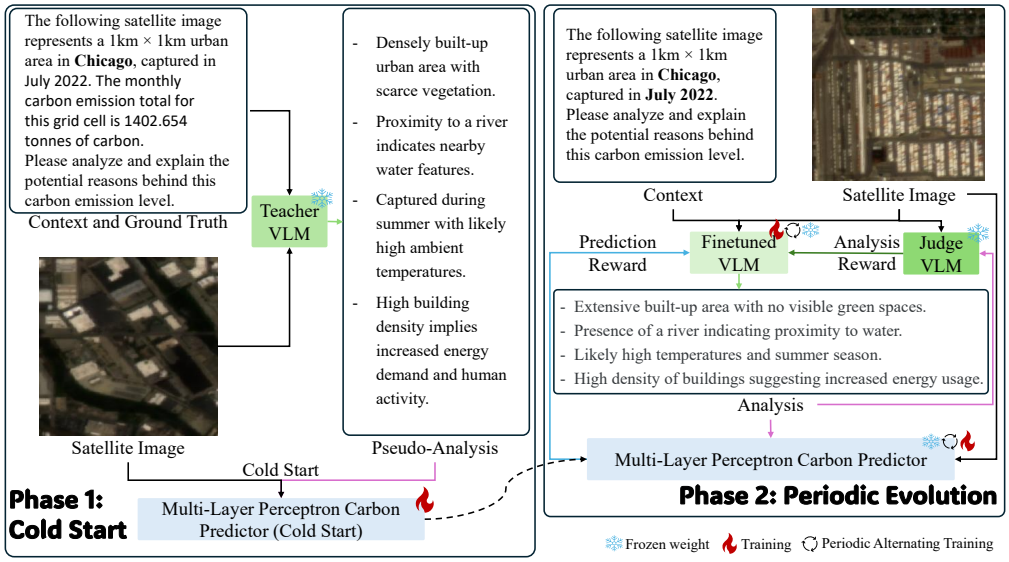


Figure 2: Training framework of TRACE introduced in Fig. 1(c). Phase 1 initializes the Predictor; Phase 2 alternates Analyst and Predictor updates. The frozen visual and text encoders feeding the MLP are omitted for clarity; the visual encoder can be selected from the evaluated backbones.

We use Qwen2.5-VL-72B (Bai et al., 2025b) as the cold-start Teacher. For each sample in the Predictor-training subset, the Teacher receives the satellite image I, city, month, and the ODIAC-derived reference emission value z. Providing z guides the Teacher to generate a more targeted and emission-relevant analysis than an unconstrained image caption. The resulting pseudo-analyses are used only to initialize the Predictor f_θ. Appendix A.1 provides the prompt.

After initialization, Teacher-generated pseudo-analyses are replaced by analyses from the evolving Analyst. Predictor updates reuse the Predictor-training subset; Analyst rewards are computed on the spatially disjoint Analyst-training subset. Thus, the frozen Predictor evaluates candidate analyses on locations outside its own training subset. The labels supplied to the Teacher never come from Analyst-training samples. Validation and test samples remain excluded from both training stages.

The resulting initialization set is $\{(I_i, T_{pseudo,i}, y_i)\}_{i=1}^{N}$. The Predictor f_θ fuses visual features from I_i with a text embedding of $T_{pseudo,i}$ and outputs a continuous estimate $\hat{y}_i$. We initialize f_θ by minimizing mean squared error (MSE):

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N} \|y_i - f_\theta(I_i, T_{pseudo,i})\|^2 \tag{1}$$

This stage initializes the Predictor used to compute the Prediction Reward before Analyst RL begins.

3.3 Periodical Evolution

After cold-start initialization, we optimize the Analyst as a policy over analysis texts using the Analyst-training subset. Given an image and its context, the policy generates complete analyses that can relate multiple visual cues to emissions. Without explicit constraints, however, these analyses may contain visually unsupported content that appears relevant to emissions, or accurately describe the image without identifying cues useful for emission estimation. We therefore design complementary rewards to encourage visual grounding and predictive utility.

To keep the reward target stable within each Analyst stage, Periodical Evolution freezes the Predictor during Analyst optimization and updates it between stages using regenerated analyses. Section 3.3.3 specifies the alternating updates.

3.3.1 REWARD FUNCTION

The reward function consists of two complementary components: an Analysis Reward and a Prediction Reward. The Analysis Reward uses a Judge-based Analysis Score (AS) to evaluate grounding, carbon relevance, and logical consistency, while the Prediction Reward measures the reduction in regression error when analysis text is added to image features.

Analysis Reward The Analysis Reward is the Analysis Score in $[0, 1]$ from a frozen Qwen2.5-VL-72B Judge (Bai et al., 2025b). The Judge receives the image, city and month, and generated analysis, but no emission labels. It penalizes claims that lack visual or contextual support and also checks carbon relevance and logical consistency. The prompt is in Appendix A.2.

The Teacher and Judge both use Qwen2.5-VL-72B, but the Analyst receives no supervised training on Teacher texts. The Judge and prompt are fixed across variants, and the blind human audit (Table 2) independently supports the observed quality gains.

Prediction Reward The Prediction Reward compares the relative errors of the Predictor's image-branch and fused predictions. Let y denote the reference carbon emission, e_{img} the relative error from visual features, and e_{fused} the relative error from visual and text features:

$$e_{img} = \frac{|\hat{y}_{img} - y|}{|y| + \epsilon}, \quad e_{fused} = \frac{|\hat{y}_{fused} - y|}{|y| + \epsilon} \tag{2}$$

where $\hat{y}_{img}$ is the output of the Predictor's trained image branch, $\hat{y}_{fused}$ is its prediction from fused visual and text features, and $\epsilon = 10^{-6}$ is a small smoothing constant. We define the relative performance improvement as $\Delta = e_{img} - e_{fused}$.

Relative error accounts for differences in emission magnitude. Positive Δ indicates that analysis text improves the prediction; negative Δ indicates increased error.

For scenes that the image branch already predicts accurately, small error changes can produce a misleading reward. We therefore set $\tau = 0.05$. If $e_{img} < \tau$, the sample receives the batch mean improvement reward $\bar{R}_{batch}$; otherwise, Δ is scaled and bounded by a hyperbolic tangent:

$$R_{pred} = \begin{cases} \bar{R}_{batch}, & \text{if } e_{img} < \tau \\ \tanh(n\Delta), & \text{otherwise} \end{cases} \tag{3}$$

where n controls sensitivity to the improvement. The reward therefore favors text that adds information on samples the image branch does not already solve.

When e_{img} is below 5%, small numerical fluctuations can dominate the remaining error difference and provide little guidance about analysis content. Assigning the batch mean avoids treating these samples as large successes or failures. For other samples, $\tanh(\cdot)$ preserves the sign of Δ while bounding extreme rewards from difficult tiles.

3.3.2 UPDATE STRATEGY

We optimize the Analyst with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a critic-free variant of PPO (Schulman et al., 2017). For each image–context query, GRPO samples multiple candidate analyses and computes group-relative advantages from their composite rewards. Ranking candidates within the same query controls for sample difficulty, while a clipped policy update and KL regularization stabilize training and discourage language that exploits the current Predictor. The complete objective is provided in Appendix C.

3.3.3 ALTERNATING EVOLUTION

Jointly optimizing the Analyst VLM and predictor can result in moving-target instability. We therefore alternate discrete stages in which one module is optimized while the other remains frozen.

Analyst VLM Optimization With Predictor f_θ frozen, the Analyst VLM samples analyses given image I and context C (city and month). GRPO maximizes the expected composite reward:

$$\max_{\phi} \ \mathbb{E}_{T\sim\pi_\phi(\cdot|I,C)}\big[R(I,C,T,y)\big], \tag{4}$$

where R combines both rewards under the current predictor.

Predictor Evolution After each Analyst stage, we freeze ϕ and run π_ϕ over the Predictor-training subset to generate analyses $\{T_{evolved}\}$. The Predictor f_θ is fine-tuned on the updated image and analysis pairs (I, T_{evolved}) and labels y:

$$\min_{\theta} \ \mathbb{E}\left[\|f_\theta(I, T_{\text{evolved}}) - y\|_2^2\right]. \tag{5}$$

4 Experiments

4.1 Experiment Setup

4.1.1 Dataset

Our dataset covers five developed metropolitan cities: Beijing, Chicago, New York, Shanghai, and Singapore. All satellite imagery and carbon emission data correspond to the year 2022, ensuring temporal alignment across modalities.

Satellite Imagery We construct monthly observations from Planet imagery at 3 m per pixel (Planet, 2024). For each city and month, we select one scene after filtering candidates to less than 10% cloud cover. Automated cloud masks and manual inspection of RGB composites remove scenes with residual contamination. This retains monthly coverage while limiting atmospheric artifacts.

Carbon Emission Data Monthly CO_2 emissions, expressed as carbon mass, come from the Open-source Data Inventory for Anthropogenic CO_2 (ODIAC) (Oda et al., 2018; 2023). ODIAC provides globally consistent fossil fuel CO_2 emissions with a spatial resolution of up to 1km and monthly temporal frequency, which has been widely adopted in urban-scale carbon monitoring, emissions quantification, and policy evaluation (Yan et al., 2024; Mun & Jung, 2025; Meng et al., 2026).

Image-Label Pair Construction We crop satellite imagery to the ODIAC grid (nominally 1 km) and pair each patch with its monthly emission total. See Appendix E for data and metric definitions.

Evaluation Split For each city, we assign spatial tiles to Predictor-training, Analyst-training, validation, and test subsets in a 4:4:1:1 ratio. All available monthly samples from a tile remain in the same subset. The Predictor-training and Analyst-training subsets are spatially disjoint. The Predictor-training subset is used for both cold-start initialization and every subsequent Predictor update, whereas the Analyst-training subset is used only for RL optimization of the Analyst.

We split by tile rather than by monthly image. A random month-level split could place the same physical location in different subsets, allowing persistent road and building patterns to cross the data boundary. Keeping all months from one tile together removes this direct spatial overlap. Separating the two training subsets ensures that the Prediction Reward used to optimize the Analyst is always evaluated on locations not used to fit the Predictor. Validation and test tiles remain excluded from both optimization roles. This protocol evaluates generalization to unseen locations within each city.

4.1.2 Baselines and Evaluation

Our main comparison evaluates four image-only baselines: ViT-L/16 pre-trained on ImageNet-21k (Dosovitskiy et al., 2021; Deng et al., 2009), RemoteCLIP (Liu et al., 2024a), GeoRSCLIP (Zhang et al., 2024), and DINOv3-sat-493M (Siméoni et al., 2025). Each baseline predicts carbon emissions from its visual features without generated analysis text. We pair each backbone with TRACE-4B to evaluate whether the benefit of prediction-aware analysis persists across visual representations. Each Overall result uses one Predictor fitted on pooled five-city data; appendix city rows use separately fitted city-specific Predictors. All variants share the 4:4:1:1 tile-level split.

We evaluate pooled results from all five cities using the coefficient of determination (R^2), normalized root mean squared error (NRMSE), and mean absolute percentage error (MAPE). All three are

Backbone	$R^2 \uparrow$			NRMSE $\downarrow$			MAPE $\downarrow$		
	Base	+ TRACE	Rel.	Base	+ TRACE	Rel.	Base	+ TRACE	Rel.
ViT	0.7291	**0.7789**	+6.8%	0.0686	**0.0620**	-9.7%	0.4828	**0.3303**	-31.6%
RemoteCLIP	0.7512	**0.7955**	+5.9%	0.0658	**0.0596**	-9.3%	0.3887	**0.3438**	-11.6%
GeoRSCLIP	0.7731	**0.8171**	+5.7%	0.0628	**0.0564**	-10.2%	0.3732	**0.2880**	-22.8%
DINOv3-sat-493M	0.8364	**0.8662**	+3.6%	0.0533	**0.0482**	-9.6%	0.3303	**0.2608**	-21.0%

Table 1: Overall performance across visual backbones using data from all five cities under the same 4:4:1:1 tile-level split. Base: visual backbone without analysis text; + TRACE: the same backbone with the TRACE framework and a 4B Analyst. Relative change $\mathrm{Rel.} = (\mathrm{TRACE} - \mathrm{Base})/\mathrm{Base}$; positive is better for R^2, negative for NRMSE and MAPE. **Bold** marks the better value in each pair.

dimensionless; MAPE is reported as a fraction. Higher R^2 and lower errors are better. Table 1 pairs image-only baselines with TRACE-4B and reports relative changes as percentages.

Additional comparisons with ViT and DINOv3-sat use analyses from Qwen3-VL-2B and Qwen3-VL-4B without task-specific fine-tuning as text inputs to the Predictor. We compare these variants with TRACE-2B and TRACE-4B, where the suffix denotes Analyst size, to examine the effect of Analyst training and model scale. These comparisons and the corresponding city-level results appear in Appendices B.1 and B.2; Appendix B.3 additionally evaluates Qwen2.5-VL-3B as the Analyst.

4.1.3 Implementation Details

We use Qwen3-VL-4B (Bai et al., 2025a) as the Analyst VLM for the main prediction experiments. The Predictor encodes the satellite image with the selected visual backbone and the generated analysis with Qwen3-Embedding-0.6B (Zhang et al., 2025), which produces a 1024-dimensional text embedding. The two embeddings are concatenated and passed to an MLP that outputs the continuous carbon-emission estimate. ViT-L/16 uses a 1024-dimensional image embedding; the other configurations use features from their respective encoders.

We implement Periodical Evolution using the distributed rollout and policy-optimization components of EasyR1 (Sheng et al., 2025; Zheng et al., 2025). The Teacher is used only to generate pseudo-analyses for cold-start initialization, while the Judge is used only to compute the Analysis Reward and the Analysis Score during evaluation. Neither model is required at inference.

4.2 Main Results

4.2.1 Prediction Performance

Table 1 shows that TRACE-4B improves all three metrics over each of the four image-only baselines. Relative R^2 gains range from 3.6% to 6.8%, while NRMSE and MAPE decrease by 9.3%–10.2% and 11.6%–31.6%, respectively. All twelve backbone–metric pairs improve on the pooled test set.

With ViT, TRACE-4B raises R^2 from 0.7291 to 0.7789 and reduces NRMSE from 0.0686 to 0.0620 and MAPE from 0.4828 to 0.3303. The gains persist with stronger image-only baselines: RemoteCLIP and GeoRSCLIP achieve relative R^2 improvements of 5.9% and 5.7%, accompanied by MAPE reductions of 11.6% and 22.8%. DINOv3-sat-493M yields the best absolute performance among the tested configurations, reaching R^2 of 0.8662, NRMSE of 0.0482, and MAPE of 0.2608 with TRACE-4B. Even for this strongest image-only baseline, NRMSE and MAPE decrease by 9.6% and 21.0%, indicating that generated analyses complement the visual features for prediction.

The gains span both general-purpose and remote-sensing visual representations. RemoteCLIP, GeoRSCLIP, and DINOv3-sat already outperform the image-only ViT baseline, yet each benefits from the analysis branch. The accompanying reductions in MAPE show that the improvement extends to errors measured relative to individual targets, alongside the squared-error improvements captured by R^2 and NRMSE. This pattern supports using TRACE to augment an existing visual representation while preserving the choice of encoder for the application.

Model	AS ↑	Factuality ↑	Relevance ↑	H.-free ↑
Qwen3-VL-4B	0.6875	6.14	8.04	6.90
TRACE-4B	**0.7556**	**6.64**	**8.66**	**7.34**

Table 2: 4B analysis quality: Judge AS and human means (2 evaluators, 50 cases).

	Generated Analysis
	**Qwen3-VL-2B** - Urban density and grid layout suggest a compact city with *high building concentration*, common in densely populated regions like Shanghai, which supports significant vehicular and energy-related emissions. - A visible urban planning pattern with dense residential and commercial buildings, including **large-scale housing and commercial structures**, indicates frequent energy consumption for heating, cooling, and lighting, especially during the warmer season. - The presence of both *green spaces* and built-up zones suggests a mixed-use urban environment where outdoor emissions are balanced by some green infrastructure.
	TRACE-2B - The urban layout in Shanghai, as shown in the 1km × 1km image, features a high-density residential grid with *numerous buildings and narrow streets*. This configuration is consistent with typical urban development in a major city with a strong emphasis on housing and public transit, which is known to support high levels of vehicle emissions. - The presence of *a large number of red-roofed residential buildings* indicates a high level of housing density, which is characteristic of urban areas in Shanghai, especially in the central or suburban zones. Such a dense housing stock is a major source of carbon emissions from household vehicle use, especially in the form of private cars and motorcycles. - The *visible green spaces and waterways* suggest a well-developed urban environment, but the overall *lack of visible large-scale industrial zones or major factories* implies that emissions are primarily driven by transport and energy use. This aligns with the general pattern of urban emissions in Shanghai, where transportation and energy consumption are key contributors.

Table 3: Qualitative comparison of Qwen3-VL-2B and TRACE-2B on the same image and prompt. Green italics mark visible details, red bold text marks factual errors, and blue underlining marks reasonable carbon-related inferences.

The appendix reports Analyst training, scale, and city-level comparisons. Tables A1 and A2 compare unfine-tuned Qwen3-VL analyses with TRACE under ViT and DINOv3-sat. TRACE-4B improves on TRACE-2B and the corresponding image-only baseline across all three metrics in every city.

4.2.2 Analysis Quality

We evaluate factual grounding, carbon relevance, and logical coherence with a Judge VLM. We also conduct a blind manual audit of 50 random test cases, 10 per city. Two evaluators independently view each image, its metadata, and anonymized outputs from Qwen3-VL-4B and TRACE-4B, then score factual grounding, carbon relevance, and hallucination-free quality from 1 to 10. The last criterion penalizes unsupported claims such as fabricated factories, airports, ports, or commercial districts. Appendices A.2 and A.3 provide the Judge prompt and audit criteria.

Table 2 shows that TRACE-4B improves AS from 0.6875 to 0.7556, while the blind audit exhibits the same consistent advantage across all criteria. The Judge and blind audit provide complementary assessments: one applies fixed criteria to all outputs, whereas the other tests whether readers observe the same quality difference without knowing the model identity. Table 4(a) reports a similar improvement for the 2B model, with AS increasing from 0.6504 to 0.7414.

Table 3 shows that TRACE-2B grounds its reasoning in specific visible features, red-roofed housing, narrow streets, vegetation, and waterways, and notes the absence of industrial facilities. Qwen3-VL-2B relies on generic descriptions and introduces unsupported commercial structures.

(a) Reward components (2B)

Reward strategy	R^2 ↑	AS ↑
Qwen3-VL	0.7274	0.6504
only Prediction Reward	0.7313	0.7054
only Analysis Reward	0.7379	0.7382
TRACE	**0.7686**	**0.7414**

(b) Predictor text inputs (2B)

Text input to Predictor	R^2 ↑	NRMSE ↓	MAPE ↓
Image-only ViT	0.7291	0.0686	0.4828
Empty text	0.6987	0.0724	0.5796
City/month-only text	0.7370	0.0676	0.3820
TRACE-2B analysis	**0.7686**	**0.0634**	**0.3418**

Table 4: TRACE-2B ablations with ViT-L/16: (a) rewards; (b) Predictor text inputs.

4.3 Ablation Study

Table 4 examines reward contributions (a) and text-branch information (b).

Reward Component Ablation. With ViT-L/16 and a 2B Analyst, untuned Qwen3-VL text yields lower R^2 than image-only prediction (0.7274 vs. 0.7291), motivating task-specific training. Prediction-only optimization improves analysis quality indirectly, but without direct grounding supervision it recovers little of the full model's prediction gain.

Analysis-only optimization strengthens grounding, yet its R^2 remains below the complete objective. Analysis quality alone does not ensure that the Analyst prioritizes cues useful for emission estimation. The Prediction Reward supplies this guidance; the combined objective guides the Analyst toward grounded analyses that improve prediction.

The analysis-only variant reaches $AS = 0.7382$, close to the full model's 0.7414, while its R^2 is 0.7379 compared with 0.7686. Prediction feedback therefore improves R^2 even when the AS scores are similar. Conversely, prediction-only training raises AS from 0.6504 to 0.7054, indicating that prediction feedback also encourages some of the properties assessed by the Judge. These results support both reward components.

Controlled Text-Input Ablation. Table 4(b) compares image-only prediction, an empty text branch, city/month metadata, and TRACE analyses under a controlled architecture and training protocol. Empty text adds the text encoder without semantic information and does not yield a consistent benefit, so the additional branch alone cannot explain TRACE's improvement.

City and month provide useful coarse context, but increasing R^2 from 0.7370 with only city/month metadata in the text branch to 0.7686 with TRACE analysis shows that the generated text contributes more than location and seasonality. With the image branch retained in all conditions, generated analyses improve prediction beyond image embeddings and city/month metadata.

5 Conclusion and Discussion

TRACE provides a textual reasoning-augmented framework for carbon-emission estimation, coupling VLM-generated environmental analyses with visual features and a numerical Predictor. Evaluations across five cities and several visual backbones show improvements over image-only models and counterparts supplied with analyses from VLMs without task-specific fine-tuning. Across the four evaluated visual backbones, relative R^2 improvements over image-only baselines range from 3.6% to 6.8%, with reductions in both NRMSE and MAPE. Ablations with ViT-L/16 and the 2B Analyst indicate that these gains arise from prediction-aware analytical content rather than the additional text branch or city/month metadata alone, with the two rewards contributing complementary supervision. The resulting analyses are more grounded and carbon-relevant. Readers can inspect the stated visual evidence alongside each estimate.

Limitations Evaluation covers unseen tiles within five cities; broader geographic transfer remains a direction for future work. Generated analyses provide inspectable evidence, while causal attribution and the Predictor's reliance on individual claims require further study. Other priorities include calibrated uncertainty and region-level grounding. Cold-start initialization depends on Teacher quality, and Periodical Evolution adds training cost.

AI USE STATEMENT

We used OpenAI Codex for language editing, manuscript restructuring and formatting, figure-label revisions, coding assistance and debugging, and assistance with experimental checks and interpretation of results. The authors reviewed the AI-assisted text and ran checks on the assisted code. The Teacher, Analyst, and Judge VLMs are components of the research method, with their roles described in the methodology and appendix. The authors take responsibility for the final text, claims, code, and reported results, including material produced with AI assistance.

REFERENCES

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Dash: Detection and assessment of systematic hallucinations of vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22748–22759, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477, 2024.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*, 2025.

Monica Crippa, Diego Guizzardi, Enrico Pisoni, Efisio Solazzo, Antoine Guion, Marilena Muntean, Aneta Florczyk, Marcello Schiavina, Michele Melchiorri, and Andres Fuentes Hutfilter. Global anthropogenic emissions in urban areas: patterns, trends, and challenges. *Environmental Research Letters*, 16(7):074033, jul 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Ji Han, Xing Meng, Hanwei Liang, Zhi Cao, Liang Dong, and Cheng Huang. An improved nightlight-based method for modeling urban co2 emissions. *Environmental Modelling & Software*, 107: 307–320, 2018.

Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. Urban-vlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 28061–28069, 2025.

IPCC. Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels. Special report, Intergovernmental Panel on Climate Change, Geneva, Switzerland, 2018. URL https://www.ipcc.ch/sr15/.

Aisha Javed, Taeheon Kim, Changhui Lee, Jaehong Oh, and Youkyung Han. Deep learning-based detection of urban forest cover change along with overall urban changes using very-high-resolution satellite images. *Remote Sensing*, 15(17):4285, 2023.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Franck Lespinas, Yilong Wang, Grégoire Broquet, François-Marie Bréon, Michael Buchwitz, Maximilian Reuter, Yasjka Meijer, Armin Loescher, Greet Janssens-Maenhout, Bo Zheng, et al. The potential of a constellation of low earth orbit satellite imagers to monitor worldwide fossil fuel co2 emissions from large cities and point sources. *Carbon Balance and Management*, 15(1):18, 2020.

Haoyang Li, Xuejia Chen, Zhanchao Xu, Darian Li, Nicole Hu, Fei Teng, Yiming Li, Luyu Qiu, Chen Jason Zhang, Li Qing, et al. Exposing numeracy gaps: A benchmark to evaluate fundamental numerical abilities in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20004–20026, 2025.

Jian Lin, Qin Ma, Yang Ju, Hongsheng Zhang, Qiang Wang, and Bo Huang. Relationships between urbanization, tree morphology, and carbon density: An integration of remote sensing, allometric models, and field survey. *Urban Forestry & Urban Greening*, 76:127725, 2022.

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, 2024a.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024b.

Tianhui Liu et al. Cityrise: Reasoning urban socio-economic status in large vision-language models via reinforcement learning. *arXiv preprint arXiv:2510.22282*, 2025.

Xiaoqian Liu, Chen You, Javier Cifuentes-Faura, and Xiaoyan Peng. Bigger cities and less carbon? government-driven urbanization and carbon emissions in china. *Journal of Cleaner Production*, 467:142808, 2024c.

Mohammad Marjani, Masoud Mahdianpari, Daniel J Varon, and Fariba Mohammadimanesh. The integration of vision transformers and sam for automated methane super-emitter detection using tropomi data. *Journal of Environmental Management*, 393:127034, 2025.

Jijun Meng, Can Wang, Yi Zou, Ziyan Han, and Guannan Zou. Multi-scale spatiotemporal interactions between land use transformation and carbon emissions in china from 1980 to 2020. *Resources, Conservation and Recycling*, 226:108653, 2026.

Hansol Mun and Juchul Jung. Impact of compact city on carbon emission reduction based on urban size: A spatial analysis using satellite imagery. *Sustainable Cities and Society*, 126:106326, 2025.

Tomohiro Oda, Shamil Maksyutov, and Robert J Andres. The open-source data inventory for anthropogenic co 2, version 2016 (odiac2016): a global monthly fossil fuel co 2 gridded emissions data product for tracer transport simulations and surface flux inversions. *Earth System Science Data*, 10(1):87–107, 2018.

Tomohiro Oda, R Bun, V Kinakh, P Topylko, M Halushchak, G Marland, T Lauvaux, M Jonas, and R Zhuravlev. The open-source data inventory for anthropogenic co2 (odiac) 2023, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Planet. Planet application program interface: In space for life on earth. `https://api.planet.com`, 2024. San Francisco, CA.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

David P Roy, Haiyan Huang, Rasmus Houborg, and Vitor S Martins. A global analysis of the temporal availability of planetscope high spatial resolution multi-spectral imagery. *Remote Sensing of Environment*, 264:112586, 2021.

Muhamad Risqi U Saputra, Irfan Dwiki Bhaswara, Bahrul Ilmi Nasution, Michelle Ang Li Ern, Nur Laily Romadhotul Husna, Tahjudil Witra, Vicky Feliren, John R Owen, Deanna Kemp, and Alex M Lechner. Multi-modal deep learning approaches to semantic segmentation of mining footprints with multispectral satellite imagery. *Remote Sensing of Environment*, 318:114584, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yu-Sheng Shen, Ying-Chen Lin, Shenghui Cui, Yanmin Li, and Xingxing Zhai. Crucial factors of the built environment for mitigating carbon emissions. *Science of The Total Environment*, 806:150864, 2022. ISSN 0048-9697.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL `https://arxiv.org/abs/2508.10104`.

Zehua Song, Xiaoyang Yu, Yu Song, Jin Yang, Dongsheng Xu, Kejin Chen, Fangfei Huang, Bin Chen, and Yanwei Song. A real-time inversion framework for carbon equivalent emissions in oil and gas extraction based on vision transformer. In *ISOPE International Ocean and Polar Engineering Conference*, pp. ISOPE–I. ISOPE, 2024.

UNFCCC. The paris agreement, 2015. URL `https://unfccc.int/sites/default/files/english_paris_agreement.pdf`. Accessed: 2024-05-20.

Swati Uniyal, Saurabh Purohit, Kuldeep Chaurasia, Sitiraju Srinivas Rao, and Eadara Amminedu. Quantification of carbon sequestration by urban forest using landsat 8 oli and machine learning algorithms in jodhpur, india. *Urban forestry & urban greening*, 67:127445, 2022.

Qiongyan Wang et al. Urban-r1: Reinforced mllms mitigate geospatial biases for urban general intelligence. *arXiv preprint arXiv:2510.16555*, 2025.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. Self-rewarding correction for mathematical reasoning. *arXiv preprint arXiv:2502.19613*, 2025.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*, pp. 4006–4017, 2024.

Di Yang, Weixin Luan, Lu Qiao, and Mahardhika Pratama. Modeling and spatio-temporal analysis of city-level carbon emissions based on nighttime light satellite imagery. *Applied Energy*, 268: 114696, 2020.

Yueting Yang, Xintong Zhang, Jinan Xu, and Wenjuan Han. Empowering vision-language models for reasoning ability through large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10056–10060. IEEE, 2024.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 57654–57689. Curran Associates, Inc., 2025. doi: 10.52202/085713-1933.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025.

Kaiwen Zhou, Kwonjoon Lee, Teruhisa Misu, and Xin Wang. Vicor: Bridging visual understanding and commonsense reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10783–10795, 2024.

Shaoqing Zhou, Xiaoman Zhang, Shiwei Chu, Tiantian Zhang, and Junfei Wang. Research on remote sensing image carbon emission monitoring based on deep learning. *Signal Processing*, 207:108943, 2023.

APPENDIX: SUPPLEMENTARY MATERIAL

A PROMPT TEMPLATES

A.1 PSEUDO-ANALYSIS GENERATION PROMPT

Cold-start training does not use expert-written analysis labels. For each sample in the Predictor-training subset, a Teacher VLM receives the satellite image, city, month, and ODIAC-derived reference emission value and generates a pseudo-analysis. These pseudo-analyses are used only to initialize the Predictor and are then discarded, while the same subset is retained for subsequent Predictor updates. The Analyst is optimized on a spatially disjoint Analyst-training subset. The reference emission is provided to the Teacher as an input only during cold start; in later stages, emission labels are used only as training targets or for reward computation. The prompt follows:

> The following satellite image represents a 1km $\times$ 1km urban area in {city}, captured in {month} 2022. The total carbon emission of this grid cell during the month is $\{z\}$ tonnes of carbon.
>
> Analyze and explain the plausible and evidence-based reasons behind this observed carbon emission level. Your explanation should:
>
> 1. Be grounded in realistic reasoning based on typical urban patterns, geography, climate, and socio-economic context of {city} (avoid fictional or speculative claims).
> 2. Consider both direct emission sources (e.g., traffic, industry, energy use) and indirect influencing factors (e.g., urban planning, building density, cultural habits, policy, or seasonal energy demand).
> 3. Include quantitative or contextual justification where possible.
> 4. Present the analysis in a structured, bullet-point format, with each point representing a distinct contributing factor.

A.2 ANALYSIS QUALITY SCORE PROMPT

The Analysis Score comes from a frozen Qwen2.5-VL-72B Judge. Given the image, city, month, and generated analysis, it returns a score from 0 to 1 based on visual grounding, carbon relevance, and logical consistency. We use the same prompt for every evaluated output:

> You are a strict expert grader in environmental science. Criteria for scoring (0 to 1):
>
> - Visual grounding: Claims about objects, land use, or infrastructure must be supported by the image. Fabricated or visually unsupported scene content should substantially reduce the score.
> - Carbon relevance: The analysis should connect supported evidence to carbon-emission estimation rather than list generic environmental factors. It is not required to mention factors that are absent from the image.
> - Logical consistency: The explanation should remain internally coherent and should not draw conclusions that are unsupported by its stated evidence.
>
> Only output the final score as a float inside ¡score¿...¡/score¿ tags.
>
> Example output format: ¡score¿0.85¡/score¿
>
> Evaluate the following analysis about carbon emission estimation based on the provided satellite image, which represents a 1km $\times$ 1km urban area in {city}, captured in {month} 2022.
>
> Analysis from another analyst: {analysis text}

A.3 ANALYSIS QUALITY EVALUATION PROTOCOL

For automatic evaluation, the Judge VLM applies the prompt in Appendix A.2 to the satellite image, city, month, and generated analysis. It returns a scalar in $[0, 1]$. The Judge and its prompt remain fixed across every output from the 2B and 4B variants.

The blind human audit contains 50 randomly sampled test cases, 10 from each city. Two evaluators independently inspect the image, metadata, and anonymized outputs from Qwen3-VL-4B and TRACE-4B, then assign separate scores from 1 to 10 for factuality, carbon relevance, and hallucination-free

City	Metric	Image-only	Qwen3-VL-2B	Qwen3-VL-4B	TRACE-2B	TRACE-4B
Beijing	R^2 (↑)	0.4033	0.4228	0.4224	<u>0.4797</u>	**0.5083**
	NRMSE (↓)	0.1250	0.1229	0.1230	<u>0.1167</u>	**0.1135**
	MAPE (↓)	0.3489	0.3179	0.3574	<u>0.2862</u>	**0.2806**
Chicago	R^2 (↑)	0.3951	0.4183	0.4296	<u>0.4524</u>	**0.4727**
	NRMSE (↓)	0.1246	0.1221	0.1209	<u>0.1185</u>	**0.1163**
	MAPE (↓)	0.2430	0.2421	0.2437	<u>0.2414</u>	**0.2356**
New York	R^2 (↑)	0.0928	0.0987	0.1143	<u>0.1332</u>	**0.1397**
	NRMSE (↓)	0.0757	0.0754	0.0747	<u>0.0739</u>	**0.0737**
	MAPE (↓)	0.4122	0.3907	0.4090	<u>0.3756</u>	**0.3630**
Shanghai	R^2 (↑)	0.6146	0.6284	0.6548	<u>0.6734</u>	**0.6918**
	NRMSE (↓)	0.1034	0.1015	0.0978	<u>0.0952</u>	**0.0924**
	MAPE (↓)	0.3575	0.3286	0.2895	<u>0.2787</u>	**0.2672**
Singapore	R^2 (↑)	0.4062	0.4164	0.4297	<u>0.4832</u>	**0.5112**
	NRMSE (↓)	0.1530	0.1517	0.1499	<u>0.1427</u>	**0.1388**
	MAPE (↓)	0.4522	<u>0.4073</u>	0.4173	0.4084	**0.3873**
Overall	R^2 (↑)	0.7291	0.7274	0.7425	<u>0.7686</u>	**0.7789**
	NRMSE (↓)	0.0686	0.0688	0.0669	<u>0.0634</u>	**0.0620**
	MAPE (↓)	0.4828	0.4730	0.3718	<u>0.3418</u>	**0.3303**

Table A1: Analyst training and scale comparisons with ViT under the 4:4:1:1 tile-level split. Qwen3-VL columns use analyses without task-specific fine-tuning; TRACE columns use trained Analysts. City rows use separately fitted city-specific Predictors; Overall uses a Predictor fitted on pooled five-city data. Higher R^2 and lower NRMSE/MAPE are better. **Bold**/<u>underlined</u>: best/second-best.

quality. For each criterion, we take the arithmetic mean of the two evaluators' scores for each case, then average over the 50 cases. These criteria distinguish incorrect scene descriptions from irrelevant discussion and claims about objects or land uses that are not visible. The audit is limited to the 4B pair, whereas the automatic Judge covers both model scales.

B Additional Experimental Results

B.1 Analyst Training and Scale with ViT

Table A1 compares image-only ViT, Predictors using analyses from Qwen3-VL without task-specific fine-tuning, and TRACE at the 2B and 4B Analyst scales. It complements the overall backbone comparison in Table 1 by reporting results for each city and isolating comparisons between Analyst variants within the ViT configuration.

TRACE-4B achieves the best values for all three metrics in each city and overall. TRACE-2B outperforms Qwen3-VL-2B in 17 of 18 comparisons (the exception is Singapore MAPE, 0.4084 vs. 0.4073). TRACE-4B outperforms TRACE-2B across all reported comparisons.

B.2 Analyst Training and Scale with DINOv3

Table A2 repeats the Analyst training and scale comparisons with DINOv3-sat-493M (Siméoni et al., 2025) as the visual backbone. TRACE-4B achieves the best values in every row.

The city-level DINOv3 results appear in Table A2. TRACE-4B improves all three metrics in Beijing, Chicago, New York, Shanghai, and Singapore. The absolute R^2 increase is 0.0505 in Beijing and 0.0406 in New York, compared with 0.0211 in Shanghai, where DINOv3-only already reaches 0.8173. NRMSE and MAPE follow the same direction in every city.

City	Metric	Image-only	Qwen3-VL-2B	Qwen3-VL-4B	TRACE-2B	TRACE-4B
Beijing	R^2 (↑)	0.5985	0.6249	0.6264	<u>0.6406</u>	**0.6490**
	NRMSE (↓)	0.1025	0.0991	0.0989	<u>0.0970</u>	**0.0959**
	MAPE (↓)	0.2811	0.2475	0.2437	<u>0.2427</u>	**0.2403**
Chicago	R^2 (↑)	0.6750	0.6868	0.6928	<u>0.7000</u>	**0.7066**
	NRMSE (↓)	0.0913	0.0896	0.0888	<u>0.0877</u>	**0.0867**
	MAPE (↓)	0.1737	0.1775	0.1786	<u>0.1710</u>	**0.1672**
New York	R^2 (↑)	0.2106	0.2276	0.2261	<u>0.2408</u>	**0.2512**
	NRMSE (↓)	0.0706	0.0698	0.0699	<u>0.0692</u>	**0.0687**
	MAPE (↓)	0.2933	0.3166	0.2932	<u>0.2808</u>	**0.2799**
Shanghai	R^2 (↑)	0.8173	0.8189	0.8312	<u>0.8370</u>	**0.8384**
	NRMSE (↓)	0.0712	0.0709	0.0684	<u>0.0672</u>	**0.0669**
	MAPE (↓)	0.2491	0.2368	0.2420	<u>0.2196</u>	**0.2184**
Singapore	R^2 (↑)	0.5886	0.5939	0.5981	<u>0.6078</u>	**0.6162**
	NRMSE (↓)	0.1273	0.1265	0.1259	<u>0.1243</u>	**0.1230**
	MAPE (↓)	0.3872	0.3889	0.3794	<u>0.3788</u>	**0.3742**
Overall	R^2 (↑)	0.8364	0.8387	0.8533	<u>0.8613</u>	**0.8662**
	NRMSE (↓)	0.0533	0.0529	0.0505	<u>0.0491</u>	**0.0482**
	MAPE (↓)	0.3303	0.3059	0.2772	<u>0.2654</u>	**0.2608**

Table A2: Analyst training and scale comparisons with DINOv3-sat-493M under the same protocol and notation as Table A1. Qwen3-VL columns use analyses without task-specific fine-tuning; TRACE columns use trained Analysts. City rows use separately fitted city-specific Predictors; Overall uses a Predictor fitted on pooled five-city data. **Bold**/<u>underlined</u>: best/second-best.

B.3 Alternative Analyst: Qwen2.5-VL-3B

To examine Analyst dependence, we replace Qwen3-VL with Qwen2.5-VL-3B (Bai et al., 2025b) while retaining the training procedure. These experiments cover two Qwen-VL generations.

With Qwen2.5-VL-3B as the Analyst, overall R^2 rises from 0.7291 to 0.7705 with ViT and from 0.8364 to 0.8616 with DINOv3; both error metrics decrease. Every city and both backbones follow the same direction in Table A3. Together with the Qwen3-VL results, the evidence covers two Analyst generations and two visual feature regimes within the tested model families.

C GRPO Optimization Details

TRACE uses Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a critic-free variant of PPO (Schulman et al., 2017), to update the Analyst VLM. For each query q, we sample a group of candidate analyses $\{o_1, \ldots, o_G\}$ from the old policy $\pi_{\phi_{\text{old}}}$, compute the composite reward R_i for each candidate, and derive its group-relative advantage $\hat{A}_i$. The updated policy π_ϕ maximizes

$$J_{\text{GRPO}}(\phi) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\mathcal{L}_i \;-\; \beta\, D_{\text{KL}}(\pi_\phi \,\|\, \pi_{\text{ref}})\right], \tag{C.1}$$

where the expectation is over $q \sim P(Q)$ and $\{o_i\}_{i=1}^{G} \sim \pi_{\phi_{\text{old}}}$, and

$$\mathcal{L}_i = \min\Big(r_i(\phi)\,\hat{A}_i,\ \text{clip}(r_i(\phi),\,1-\varepsilon,\,1+\varepsilon)\,\hat{A}_i\Big), \tag{C.2}$$

with likelihood ratio

$$r_i(\phi) = \frac{\pi_\phi(o_i \mid q)}{\pi_{\phi_{\text{old}}}(o_i \mid q)}. \tag{C.3}$$

City	Metric	ViT		DINOv3-sat-493M	
		Image-only	TRACE (Qwen2.5-VL-3B)	Image-only	TRACE (Qwen2.5-VL-3B)
Beijing	R^2 (↑)	0.4033	0.4839	0.5985	0.6437
	NRMSE (↓)	0.1250	0.1163	0.1025	0.0966
	MAPE (↓)	0.3489	0.2816	0.2811	0.2469
Chicago	R^2 (↑)	0.3951	0.4563	0.6750	0.7045
	NRMSE (↓)	0.1246	0.1181	0.0913	0.0871
	MAPE (↓)	0.2430	0.2415	0.1737	0.1715
New York	R^2 (↑)	0.0928	0.1210	0.2106	0.2496
	NRMSE (↓)	0.0757	0.0745	0.0706	0.0688
	MAPE (↓)	0.4122	0.3893	0.2933	0.2813
Shanghai	R^2 (↑)	0.6146	0.6771	0.8173	0.8354
	NRMSE (↓)	0.1034	0.0946	0.0712	0.0676
	MAPE (↓)	0.3575	0.2873	0.2491	0.2247
Singapore	R^2 (↑)	0.4062	0.4699	0.5886	0.6215
	NRMSE (↓)	0.1530	0.1445	0.1273	0.1221
	MAPE (↓)	0.4522	0.4141	0.3872	0.3769
Overall	R^2 (↑)	0.7291	0.7705	0.8364	0.8616
	NRMSE (↓)	0.0686	0.0631	0.0533	0.0490
	MAPE (↓)	0.4828	0.3528	0.3303	0.2646

Table A3: TRACE with Qwen2.5-VL-3B as the Analyst VLM under the 4:4:1:1 tile-level split. City rows use separately fitted city-specific Predictors; Overall uses a Predictor fitted on pooled five-city data. The framework improves both visual backbones across all metrics.

Here, ε is the clipping threshold, β controls KL regularization with respect to the reference policy π_{ref}, and $\hat{A}_i$ is computed from the rewards of analyses sampled for the same query. The group-relative advantage therefore compares alternative explanations for one image and context rather than directly comparing samples with different emission levels or scene difficulty.

This within-query comparison makes GRPO suitable for TRACE: candidates for a difficult industrial tile are ranked against one another, as are candidates for an easier residential tile. Clipping limits abrupt policy changes, while KL regularization discourages the Analyst from moving far from the reference language distribution merely to exploit the current frozen Predictor.

D Failure Case on an Atypical Non-Urban Scene

Table A4 shows an atypical non-urban scene from Singapore in March 2022, within the study geography and year. Dense vegetation and water dominate the image, rather than the built-up structures targeted by our urban estimation task.

The model interprets circular bright structures as roundabouts or traffic islands and the adjacent water as urban water-management infrastructure. These claims lack support in the visible scene. Its air-conditioning discussion instead draws on the month metadata, rather than on vegetation. The seasonal description is imprecise: March spans the transition from the northeast monsoon to the inter-monsoon period. The response illustrates how plausible contextual reasoning can coexist with unsupported visual claims.

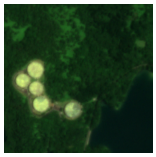
Satellite Image	Generated Analysis (TRACE-4B)
	- The image shows a cluster of circular, bright features consistent with **typical urban infrastructure such as roundabouts or circular traffic islands**, suggesting organized traffic flow typical in Singapore's urban design, which indirectly supports emission-related activity through vehicle movement. - The surrounding dense green vegetation and proximity to water suggest Singapore's high urban-green ratio, consistent with its policy-driven efforts to integrate green spaces and mitigate urban heat and emissions, though this does not directly reduce emissions from transport or energy use. - The March 2022 timing aligns with Singapore's **monsoon season**, which may influence emissions indirectly through increased rainfall affecting emissions transport (e.g., particulate dispersion) and higher energy demand for air conditioning during the inter-monsoon period, though no direct emissions are visible in this single image. - The presence of a water body adjacent to the urban cluster implies **possible urban water management infrastructure, such as reservoirs or drainage systems**, which support environmental resilience and may indirectly reduce emissions through optimized water-use efficiency in urban operations.

Table A4: Representative failure of TRACE-4B. Red bold text marks unsupported visual claims and an imprecise seasonal characterization. The image and generated analysis appear side by side.

E Data Statistics, Metric Definitions, and Label Units

E.1 Sample Counts and Spatial Separation

Table A5 reports the retained monthly image–label pairs in the split manifests. A spatial tile is identified by its city and grid row/column; all available months from a tile share one subset. The four tile sets are disjoint within every city. The 4:4:1:1 allocation applies to tiles, so monthly sample counts need not follow that ratio exactly. Not every tile has all 12 retained observations.

City	Predictor	Analyst	Validation	Test	Total	Tiles
Beijing	4,562	4,578	1,149	1,186	11,475	1,023
Chicago	3,494	3,458	896	893	8,741	846
New York	4,901	4,935	1,235	1,241	12,312	1,178
Shanghai	3,792	3,800	949	974	9,515	892
Singapore	2,476	2,526	623	646	6,271	636
Total	19,225	19,297	4,852	4,940	48,314	4,575

Table A5: Monthly sample counts by training role and evaluation subset. "Tiles" counts unique spatial locations across all four subsets.

E.2 Evaluation Metrics

For a test set of n samples, let y_i and $\hat{y}_i$ denote the target and prediction, and $\bar{y}$ the mean target. We compute

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{E.1}$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}}{\max_i y_i - \min_i y_i}, \tag{E.2}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_i \frac{|y_i - \hat{y}_i|}{\max(|y_i|, \epsilon)}. \tag{E.3}$$

The NRMSE denominator is the target range of the evaluated test set. Overall metrics are computed over the pooled test samples from all five cities; city-level metrics use the corresponding city's test set. Overall uses a Predictor fitted on pooled five-city data; city-level rows use separately fitted

city-specific Predictors. MAPE follows the scikit-learn implementation with $\epsilon = 2^{-52}$ and is reported as a fraction (e.g., 0.25 corresponds to 25%). The retained split labels are strictly positive, so the zero-denominator safeguard is inactive on these data. Extraction retains valid, positive labels.

E.3 Physical Units and Numerical Scaling

The ODIAC 1 km GeoTIFF product stores monthly emission totals in tonnes of carbon per grid cell (Oda et al., 2023). The inventory quantifies CO_2 emissions using carbon mass (C). Our extraction preserves the original raster values. The Teacher receives the raw raster value z; the regression target is $y = z/1000$, expressed in kilotonnes of carbon (kt C) per grid cell per month (Figure 1).

Consistently rescaling predictions and positive targets by the same constant leaves the three reported dimensionless metrics unchanged.