MAPDEC-R1: CARTOGRAPHIC EVIDENCE-GROUNDED REASONING FOR MAP DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated maps are becoming increasingly difficult to distinguish from authentic maps based on visual appearance alone. Unlike natural images, maps are structured representations governed by cartographic rules and spatial-semantic constraints, providing domain-specific authenticity evidence beyond the visual artifacts commonly exploited by general deepfake detectors. In this work, we propose MapDec-R1, a cartographic evidence-grounded reasoning framework for map deepfake detection. MapDec-R1 constructs a Cartographic Evidence Space that organizes authenticity cues across visual appearance, cartographic elements, and spatial-semantic relations, enabling detection based on map-specific evidence rather than generator-specific artifacts. We further introduce Cartographic Evidence-Grounded Reasoning Alignment, which aligns model reasoning with verifiable cartographic evidence instead of directly imitating teacher-generated reasoning trajectories. To jointly optimize detection and evidence-grounded reasoning, a two-stage curriculum reinforcement learning strategy first establishes reliable authenticity discrimination and then incorporates evidence-level feedback under outcome constraints. We also construct TMap-DD, a benchmark covering diverse map styles, real-map sources, and generation mechanisms. Extensive experiments demonstrate that MapDec-R1 achieves strong detection and cross-source generalization while producing interpretable authenticity reasoning grounded in cartographic evidence.

1 INTRODUCTION

With the rapid development of deep learning, large-scale generative models has reduced the barrier to map creation, allowing users with limited cartographic expertise to generate realistic geographic representations through simple instructions (Kang et al., 2024; Lin & Zhao, 2025). However, this democratization of map generation challenges the long-standing trust placed in maps as reliable sources of geographic information, as synthetic map content can be produced and disseminated without established verification processes (Prestby, 2025b;a). Therefore, developing effective methods to distinguish AI-generated maps from authentic cartographic products is becoming increasingly important for trustworthy GeoAI (Prestby, 2025a; Zhou et al., 2025; Li et al., 2025).

Existing deepfake detection methods are predominantly designed for natural images and typically distinguish synthetic content through generation artifacts, local textures, frequency patterns, or statistical discrepancies (Huang et al., 2025a; Gao et al., 2024; Cai et al., 2025). Such cues alone are insufficient for maps because they primarily capture how an image is generated rather than whether it conforms to cartographic principles (Zhao et al., 2021). Unlike natural images, maps are rule-governed visual representations constructed through cartographic abstraction and symbolization, where the organization and relationships among map elements serve as essential indicators of authenticity, as illustrated in Figure 1 (Prestby, 2025b). Moreover, low-level generation artifacts can vary substantially across generators and map styles, limiting their ability to provide stable evidence for unseen sources (Ojha et al., 2023; Wu et al., 2026). Robust map deepfake detection should therefore move beyond identifying what looks synthetic and examine whether the observed content is consistent with underlying cartographic evidence (Meo, 2026; Li et al., 2025).

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.

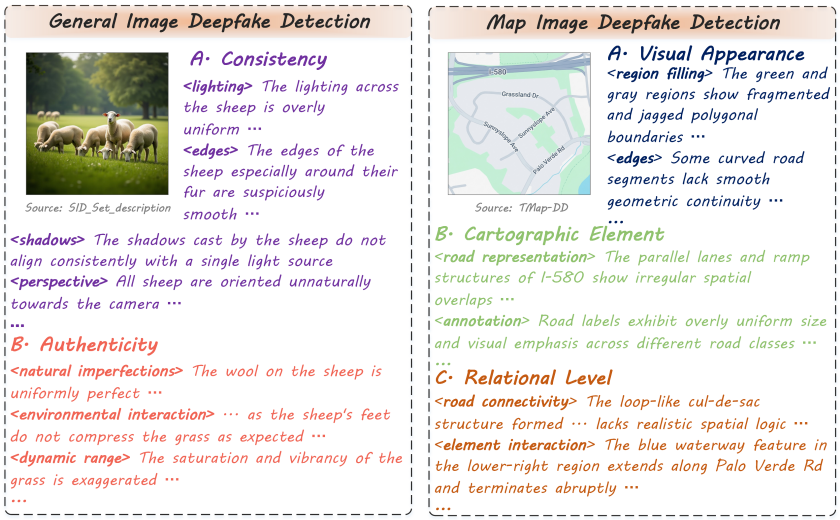

Figure 1: Comparison of deepfake detection reasoning between general images and 2D vector maps. The general image example and its corresponding forgery analysis are adapted from SID_Set_description (Huang et al., 2025b). The 2D vector map example is sampled from TMap-DD, with the corresponding cartographic forgery analysis manually curated by us based on visual, cartographic, and relational evidence.

Multimodal large language models (MLLMs) provide a promising foundation for this goal because they can jointly interpret visual content and generate semantic explanations (Wu et al., 2023). However, simply introducing reasoning does not guarantee that the reasoning is relevant or faithful to map authenticity. Conventional reasoning supervision often distills teacher-generated reasoning trajectories, encouraging a model to reproduce linguistic reasoning patterns rather than verify whether its explanation is supported by observable map evidence (Hu et al., 2026). This issue is particularly important for maps, where useful authenticity cues are distributed across multiple levels, from rendering appearance and individual cartographic elements to higher-order spatial and semantic relationships. Moreover, detection and explanation provide supervision at different granularities: a binary label indicates whether the final decision is correct, whereas evidence supervision evaluates whether the reasoning itself is grounded in valid cartographic observations. Directly optimizing these objectives together can lead to conflicting learning signals.

To address these challenges, we propose MapDec-R1, a cartographic evidence-grounded reasoning framework for AI-generated map detection. Unlike conventional deepfake detection that relies on generator-specific artifacts, MapDec-R1 formulates map authenticity assessment as an evidence-grounded reasoning problem, where model decisions are supported by verifiable cartographic evidence. Based on this formulation, we introduce the Cartographic Evidence Space (CES), which organizes authenticity evidence into three complementary levels, covering visual appearance, cartographic elements, and relational consistency among map elements. Building upon CES, we further propose Cartographic Evidence-Grounded Reasoning Alignment (CEGRA), which optimizes model reasoning according to evidence consistency rather than imitation of predefined teacher reasoning trajectories. To acquire such evidence-grounded reasoning capability, we develop a two-stage curriculum reinforcement learning strategy, where the first stage establishes reliable authenticity discrimination through outcome-level supervision and the second progressively introduces evidence-level feedback under outcome constraints. We further construct TMap-DD, a benchmark for AI-generated map detection covering diverse generation mechanisms, map styles, geographic regions, and spatial scales. The training set additionally provides cartographic evidence generated under the proposed evidence space to support evidence-grounded reasoning supervision. Extensive experiments demonstrate that MapDec-R1 achieves strong detection performance across different generators and real-map sources while producing explanations grounded in concrete cartographic evidence.

Our main contributions are summarized as follows: (1) We formulate AI-generated map detection as a cartographic evidence-grounded reasoning problem, moving beyond generator-specific visual artifacts by exploiting map-specific constraints at the appearance, element, and relational levels. (2) We propose MapDec-R1, which combines Cartographic Evidence-Grounded Reasoning Alignment with two-stage curriculum reinforcement learning to jointly learn reliable authenticity discrimination and evidence-grounded explanations without directly imitating teacher reasoning trajectories. (3) We introduce TMap-DD, a benchmark covering diverse map styles, generation mechanisms, geographic regions, and spatial scales, and conduct extensive experiments to evaluate detection performance, cross-source generalization, and cartographic evidence-grounded reasoning.

2 Related Work

2.1 Multimodal Large Language Models for Geospatial Intelligence

With the rapid development of deep learning, multimodal large language models (MLLMs) have demonstrated strong capabilities in visual perception, semantic understanding, and cross-modal reasoning, enabling them to interpret complex visual content and reason over heterogeneous information (Jin et al., 2025; Wang et al., 2024). These capabilities have facilitated the increasing adoption of MLLMs in the geospatial domain, including map understanding, remote sensing image interpretation, visual geolocation, geographic question answering, and geospatial scene reasoning (Zhang et al., 2025b; 2024; Han et al., 2025). However, existing studies primarily focus on extracting and reasoning about semantic information from geospatial data, while largely overlooking the potential risks of generative manipulation within these data sources (Pisetta et al., 2026; Mai et al., 2026). In this work, we further explore the potential of MLLMs for geospatial deepfake detection by introducing a cartographic evidence-guided reasoning mechanism for explainable authenticity verification of AI-generated map tiles.

2.2 AI-Generated Image Detection

Existing studies on AI-generated content detection have primarily focused on general-purpose visual data, including images and videos (Tan et al., 2026; Huang et al., 2025a). Most approaches distinguish synthetic content by identifying generation artifacts, local texture anomalies, frequency-domain patterns, or statistical discrepancies between real and generated data (Huang et al., 2025a; Gao et al., 2024; Cai et al., 2025). However, directly applying this detection paradigm to maps remains challenging. Unlike natural images, maps are rule-governed visual representations constructed through cartographic abstraction and symbolization, where map elements are organized according to established cartographic conventions and spatial relationships. Modern generative models can readily reproduce visually plausible individual elements, such as roads, symbols, and labels, making conventional low-level visual artifacts less informative for determining map authenticity (Mahara et al., 2025; Meo, 2026). Consequently, effective detection of AI-generated maps requires moving beyond surface-level appearance and identifying cartographic evidence that reflects inconsistencies in the organization, representation, and relationships of map elements.

3 Dataset

TMap-DD-Train is constructed to capture diverse cartographic patterns and synthetic generation mechanisms for learning transferable authenticity cues. The overall composition and generation configurations of TMap-DD-Train are summarized in Table 1.

3.1 TMap-DD Training Dataset

Real Maps. We collect real map tiles from Google Maps covering diverse urban regions, including major metropolitan areas such as the San Francisco Bay Area and the Los Angeles metropolitan area. All samples are collected at zoom level 16 and manually inspected by cartography experts to ensure visual quality and cartographic consistency.

Table 1: Composition and generation configurations of TMap-DD-Train.

Category	Generation family	Generator	Samples
Real Map	–	Google Maps	623
Synthetic Map	Commercial-based	GPT-Image-1.5 (OpenAI, 2025), Nano Banana 2 (Google DeepMind, 2026b)	233
	Diffusion-based	Flux.1-dev (Black Forest Labs, 2024), Z-Image-Turbo (Z-Image Team et al., 2025), MapGenerator (Zhang et al., 2025a)	300
	GAN-based	CycleGAN (Zhu et al., 2017), SIMDCL (Xu et al., 2023), GeoMapGAN (Ma et al., 2025), SMAPGAN (Chen et al., 2020), SN-DCR (Zhao et al., 2025), Stego-GAN (Wu et al., 2024)	90
		Total	**1246**

Table 2: Distribution shifts in TMap-DD-Eval.

Evaluation axis	Training distribution	Evaluation distribution
Map source	Google Maps	Google Maps, OpenStreetMap
Zoom level	Zoom 16	Zoom 14–18
Generator	Seen generators	14 generators (7 seen + 7 unseen)

Synthetic Maps. To model diverse AI generation mechanisms, TMap-DD-Train includes synthetic map tiles generated by commercial-based, diffusion-based, and GAN-based generators with different conditioning strategies. We further incorporate publicly available GAN-based map samples from TerraFly-Forensics (Mahara et al., 2025) to expand the diversity of synthetic map distributions. This multi-generator design covers diverse forgery patterns across generation paradigms, reducing the bias toward artifacts associated with a single generator.

Cartographic Evidence Annotation. Beyond binary Beyond authenticity labels, TMap-DD provides structured cartographic evidence annotations to support fine-grained authenticity reasoning. Following the Cartographic Evidence Space, which organizes authenticity evidence into three complementary levels, including visual appearance, cartographic element representation, and inter-element relationships, we employ Gemini-3-Flash-Preview (Google DeepMind, 2025) as a teacher model to generate evidence-grounded reasoning references. Category-aware prompts are designed for real and AI-generated map tiles to guide teacher evidence discovery according to CES (see Appendix A.5). The generated evidence is further decomposed into atomic observations for fine-grained verification and supervision, followed by validation from cartography experts.

3.2 TMap-DD Evaluation Benchmark

To evaluate whether models learn transferable cartographic evidence rather than generator-specific artifacts, we construct TMap-DD-Eval with controlled distribution shifts. The evaluation protocol of TMap-DD-Eval is summarized in Table 2, with detailed configurations provided in Appendix A.1.

Real-World Distribution Shift. The evaluation benchmark contains real map tiles from multiple map sources, including Google Maps and OpenStreetMap, covering diverse geographic regions across Europe, Asia, North America, and Africa/Middle East. The samples span multiple zoom levels ranging from 14 to 18, introducing variations in geographic context and spatial scale. Compared with the fixed-source and fixed-scale training setting, these variations introduce realistic distribution shifts for evaluating out-of-distribution generalization.

Generator Shift. The synthetic evaluation set follows the same generation paradigms as the training set while introducing both seen and unseen generators. This setting evaluates generalization beyond generator-specific patterns observed during training.

4 Method

4.1 Overview

This section presents the overall framework of MapDec-R1, as illustrated in Fig. 2. We first construct a Cartographic Evidence Space (CES), which organizes authenticity evidence across visual appearance, cartographic elements, and spatial-semantic relationships. Guided by CES, teacher-generated evidence is decomposed into atomic, verifiable cartographic evidence units for subsequent

reasoning alignment. Based on this evidence space, we introduce Cartographic Evidence-Grounded Reasoning Alignment (CEGRA) and a two-stage curriculum reinforcement learning strategy to optimize evidence-grounded authenticity reasoning.

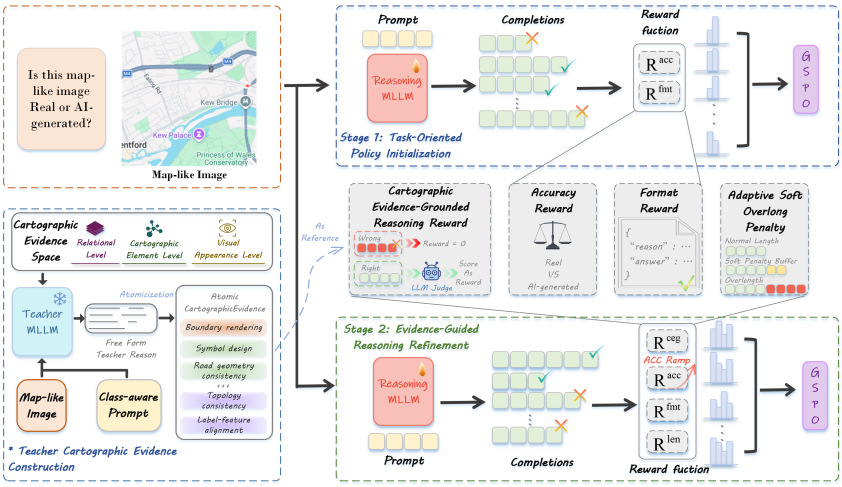

Figure 2: Overview of MapDec-R1.

4.2 Cartographic Evidence-Grounded Reasoning Alignment

MapDec-R1 formulates reasoning supervision as evidence-grounded alignment between generated explanations and structured cartographic evidence. To this end, we construct a Cartographic Evidence Space (CES) to organize authenticity evidence and guide teacher evidence discovery, and introduce Cartographic Evidence-Grounded Reasoning Alignment (CEGRA) to evaluate the semantic consistency between student reasoning and cartographic evidence.

Cartographic Evidence Space. CES organizes authenticity evidence into a three-level hierarchy, denoted as $\mathcal{E} = (\mathcal{E}^{app}, \mathcal{E}^{elem}, \mathcal{E}^{rel})$. Specifically, $\mathcal{E}^{app}$ captures rendering-related cues such as color consistency, region filling, and local textures; $\mathcal{E}^{elem}$ characterizes whether individual map elements follow cartographic representation rules; and $\mathcal{E}^{rel}$ models higher-order spatial-semantic relationships among map elements, including connectivity, semantic correspondence, and scale-dependent organization. Together, these levels constitute a hierarchical evidence space that guides the teacher model's evidence discovery process and defines the validity criteria of authenticity evidence. Given a training sample i, the teacher model produces the cartographic evidence representation $E_i = \{e_{i1}, e_{i2}, \ldots, e_{ik}\}$, where e_{ik} denotes an atomic and verifiable cartographic evidence unit.

Evidence-Grounded Alignment. Given a map tile I_i, task prompt P_i, and class label y_i, the student model samples N candidate responses from policy π_θ, $O_i^{(j)} = (R_{i,j}^S, A_{i,j}^S)$, $j = 1, \ldots, N$, where $R_{i,j}^S$ and $A_{i,j}^S$ denote the reasoning explanation and predicted label, respectively. Based on these formulations, CEGRA is defined as $S_{i,j}^{\text{CEGRA}} = J_\psi^{\text{freeze}}(I_i, P_i, y_i, E_i^T, R_{i,j}^S, A_{i,j}^S)$, where J_ψ^{freeze} denotes a frozen Judge Model that measures the semantic consistency between student reasoning and reference cartographic evidence. Through CEGRA, structured cartographic evidence is converted into evidence-grounded reasoning supervision.

4.3 Reward Functions

To support different optimization objectives in the two training stages, MapDec-R1 defines a set of reward functions that evaluate prediction correctness, evidence consistency, and reasoning quality. Different subsets of these rewards are activated according to the requirements of each training stage. For a generated response O_i, the reward set is defined as $\mathcal{R} = \{R^{\text{acc}}, R^{\text{fmt}}, R^{\text{ceg}}, R^{\text{plen}}\}$,

where R^{acc}, R^{fmt}, R^{ceg}, and R^{plen} denote the accuracy reward, Cartographic Evidence-Grounded reasoning reward, format reward, and adaptive soft overlong penalty, respectively.

Accuracy Reward. The accuracy reward provides direct supervision for authenticity prediction and is assigned as $R^{\text{acc}}_{i,j} = 1$ if the predicted label $A^s_{i,j}$ matches the ground truth y_i, and -1 otherwise.

Cartographic Evidence-Grounded Reasoning Reward. To evaluate whether generated reasoning is supported by cartographic evidence, we reuse the CEGRA score $R^{\text{ceg}}_{i,j} = S^{\text{CEGRA}}_{i,j}$, where $S^{\text{CEGRA}}_{i,j}$ measures the semantic consistency between student reasoning and reference cartographic evidence, encouraging the model to generate authenticity explanations grounded in concrete map evidence rather than generic descriptions or unsupported reasoning.

Format Reward. To ensure stable output parsing, each response is required to be a valid JSON object containing only the predefined fields. The reasoning field must be a non-empty string, and the answer field must correspond to a valid authenticity label. The reward is set to 1 if all format constraints are satisfied and 0 otherwise.

Adaptive Soft Overlong Penalty. To avoid unnecessary reasoning expansion, we introduce:

$$R^{\text{len}}_{i,j} = \min\left(-\frac{L_{i,j} - (L_{\max} - L_{\text{cache}})}{L_{\text{cache}}}, 0\right) \tag{1}$$

where $L_{i,j}$, $L_{\max}$, and L_{cache} denote the response length, maximum length threshold, and length buffer, respectively.

4.4 Two-Stage Curriculum Reinforcement Learning

In this section, we detail the two-stage curriculum reinforcement learning pipeline of MapDec-R1, including task-oriented policy initialization and evidence-guided reasoning refinement.

4.4.1 Task-Oriented Policy Initialization

In the first stage, MapDec-R1 focuses on establishing reliable authenticity prediction capability through task-oriented reinforcement learning. Given the candidate responses $O^{(j)}_i$, we activate outcome-level rewards $R^{(1)} = \lambda^{(1)}_{\text{acc}} R^{\text{acc}} + \lambda^{(1)}_{\text{fmt}} R^{\text{fmt}}$, where R^{acc} constrains authenticity label prediction, and R^{fmt} ensures that the structured output. The policy is optimized with the group-relative policy optimization objective:

$$\mathcal{J}^{(1)}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{j=1}^{G} \min\left(\rho_{i,j}(\theta)\hat{A}_{i,j}, \text{clip}(\rho_{i,j}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,j}\right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right] \tag{2}$$

where $\hat{A}_{i,j}$ denotes the relative advantage normalized within the response group of size G, and the importance ratio is

$$\rho_{i,j}(\theta) = \frac{\pi_\theta\left(O^{(j)}_i | I_i, P_i\right)}{\pi_{\theta_{\text{old}}}\left(O^{(j)}_i | I_i, P_i\right)}. \tag{3}$$

This stage focuses on answer correctness and structured prediction without introducing cartographic evidence alignment. After optimization: $\pi_{\theta_0} \rightarrow \pi_{\theta_1}$, the model obtains a stable authenticity assessment policy for subsequent evidence-guided reasoning refinement.

4.4.2 Evidence-Guided Reasoning Refinement

After outcome-level reinforcement learning, MapDec-R1 obtains a stable authenticity prediction policy. However, outcome-level feedback only supervises final decisions and does not guarantee that the prediction is grounded in valid cartographic evidence. To address this issue, we introduce outcome-conditioned evidence-grounded RL optimization, which incorporates teacher evidence into process-level policy refinement. The reward set is defined as $\mathcal{R}^{(2)} = \{R^{\text{acc}}, R^{\text{fmt}}, R^{\text{ceg}}, R^{\text{plen}}\}$.

Since outcome and evidence rewards optimize different objectives, they may conflict during policy optimization. Specifically, R^{acc} optimizes the final authenticity decision, whereas R^{ceg} evaluates

evidence attribution in the reasoning process. When the predicted outcome is incorrect, evidence rewards may reinforce locally plausible but task-misaligned reasoning patterns. Therefore, evidence supervision is restricted to outcome-consistent responses, $\bar{R}_i^{\text{ceg}} = R_i^{\text{ceg}} \mathbb{I}(R_i^{\text{acc}} > 0)$. To balance heterogeneous reward distributions, MapDec-R1 first applies within-group normalization, $z_{i,k} = (r_{i,k} - \mu_{g,k})/(\sigma_{g,k} + \epsilon)$, $k \in \{\text{acc}, \text{ceg}, \text{fmt}, \text{len}\}$. Reward normalization reduces scale discrepancies but does not determine the optimization priority among different objectives across training process. Therefore, MapDec-R1 introduces Outcome-Aware Reward Scheduling, $\hat{z}_{i,k} = \gamma_k(t) z_{i,k}$, where

$$\gamma_k(t) = \begin{cases} \gamma_s + (\gamma_e - \gamma_s) \left(\frac{t}{T}\right)^2, & k = \text{acc}, \\ 1, & \text{otherwise}. \end{cases} \tag{4}$$

where $0 \leq \gamma_s < \gamma_e$. This strategy emphasizes evidence optimization during early training to acquire cartographic reasoning patterns and gradually strengthens outcome constraints to maintain prediction reliability. Moreover, the length penalty is activated only after sufficient task competence, $\bar{R}_i^{\text{len}} = R_i^{\text{len}} \mathbb{I}\big(Acc(G_i) > \tau\big)$, where $Acc(G_i)$ denotes the average authenticity prediction accuracy within the response group. The final advantage is $A_i^{(2)} = BN\big(\sum_k \lambda_k^{(2)} \hat{z}_{i,k}\big)$, and the policy is optimized with GSPO:

$$\mathcal{J}^{(2)}(\theta) = \mathbb{E}\left[\min\left(\rho_i(\theta) A_i^{(2)}, \text{clip}(\rho_i(\theta), 1-\epsilon_l, 1+\epsilon_h) A_i^{(2)}\right)\right] - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\theta_1}) \tag{5}$$

Through Outcome-Conditioned Evidence-Grounded RL Optimization, MapDec-R1 formulates evidence-guided reasoning as a multi-objective policy optimization problem under outcome constraints, achieving dynamic coordination among heterogeneous supervision signals while learning deepfake detection strategies supported by valid cartographic evidence.

5 Experiments

5.1 Experimental Settings

Baselines. We compare MapDec-R1 with representative open-weight and proprietary VLMs across different scales and architectures. All baselines are evaluated under the same task definition, image inputs, prompts, and evaluation protocols (see Appendix A.2 for details).

Evaluation Metrics. We report Accuracy and Macro-F1 for binary authenticity classification.

Implementation Details. We use Qwen2.5-VL-3B-Instruct as the student model and perform two-stage GRPO training with MS-SWIFT 4.1. LoRA is applied only to the language model linear layers, while the vision encoder and visual alignment modules remain frozen. Both stages use sequence-level importance sampling, with one training epoch and eight sampled responses per input. Stage 2 continues from the merged Stage-1 weights and incorporates GDPO reward normalization with a frozen UnifiedReward-2.0-Qwen-7B (Wang et al., 2025b) judge model for evidence-level supervision. Additional details are provided in Appendix A.3.

5.2 Detection Performance and Generalization

5.2.1 Detection Performance

Overall Performance. As shown in Table 3, MapDec-R1 achieves the best overall performance among all evaluated VLMs, reaching 92.00% Accuracy and 91.99% Macro-F1.

Gain over the Initialization Model. Compared with Qwen2.5-VL-3B-Instruct, MapDec-R1 improves Accuracy from 45.83% to 92.00% and Macro-F1 from 35.62% to 91.99%, corresponding to gains of 46.17 and 56.37 percentage points, respectively. Since both models share the same model scale and visual configuration, this controlled comparison highlights the gains introduced by the proposed evidence-guided reinforcement learning.

Comparison with General-Purpose VLMs. Larger and newer-generation VLMs yield only limited gains on map authenticity detection. Gemini 3.8 Flash, the strongest evaluated baseline, achieves 69.33% Accuracy and 66.15% Macro-F1, while MapDec-R1 surpasses it by 22.67 and 25.84 percentage points, respectively, using a 3B-parameter backbone. These results suggest that task-specific

evidence-guided optimization provides stronger adaptation to cartographic authenticity assessment than model scaling alone.

Table 3: Performance comparison of general-purpose VLMs and MapDec-R1 on map authenticity detection.

Model	Type	Params	Acc.	Macro-F1
Open-weight VLMs				
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	Backbone	3B	0.4583	0.3562
Qwen2.5-VL-72B-Instruct (Bai et al., 2025b)	Scale-up	72B	0.5917	0.5131
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	Next-gen	4B	0.5117	0.3641
Qwen3.8-27B (Qwen Team, 2026)	Next-gen	27B	0.5767	0.4860
InternVL3.5-8B-Instruct (Wang et al., 2025a)	Cross-family	8B	0.5583	0.4919
Llama-3.2-11B-Vision-Instruct (Llama, 2023)	Cross-family	11B	0.5217	0.4164
Gemma 3-27B-IT (Team et al., 2025)	Cross-family	27B	0.6717	0.6587
Proprietary VLMs				
GPT-5.6 Sol (OpenAI, 2026)	Proprietary	–	0.5517	0.4389
Gemini 3.8 Flash (Google DeepMind, 2026a)	Proprietary	–	0.6933	0.6615
DeepSeek V4.1 Flash (Xu et al., 2026)	Proprietary	–	0.6483	0.6480
Ours				
MapDec-R1	Evidence-guided RL	3B	**0.9200**	**0.9199**

Table 4: Seen and OOD accuracy on real and generated maps.

Type	Setting	Acc.
Real	OSM (OOD)	97.50
	GoogleMap (Seen)	93.64
	Overall	**94.67**
AI-generated	OOD	88.43
	Seen	89.94
	Overall	**89.33**

Table 5: Scale generalization across zoom and map source.

Zoom	Google Maps (Seen)	OSM (OOD)	Avg.
14	97.83	100.00	98.11
15	97.73	95.00	96.88
16	98.77	100.00	99.12
17	88.46	85.71	87.88
18	65.22	100.00	77.78

5.2.2 Generalization Evaluation

Benchmark Comparison under Diverse Generation Paradigms. We evaluate MapDec-R1 across diverse map generation paradigms, including real maps and AI-generated maps from GAN, Diffusion, and GLLM generators. As shown in Fig. 3, MapDec-R1 maintains strong performance across different sources with limited performance variation, suggesting that evidence-guided reasoning provides transferable authenticity cues beyond specific generation mechanisms.

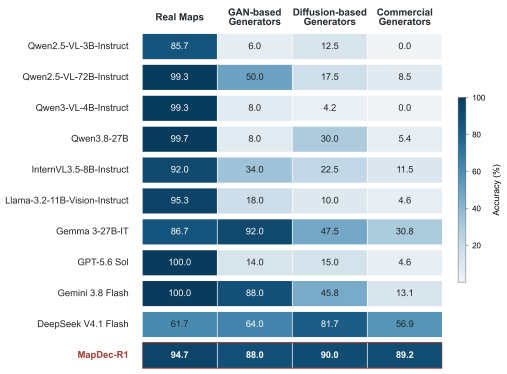

Figure 3: Cross-generator generalization performance across different map generation paradigms.

Seen and OOD Generalization Evaluation. We evaluate MapDec-R1 under seen and out-of-distribution (OOD) map sources (Table 4). For real map tiles, MapDec-R1 achieves 97.50% accuracy on the unseen OpenStreetMap source, outperforming the seen Google Maps setting (93.64%), indicating reduced dependence on provider-specific rendering patterns. For AI-generated map tiles, it maintains comparable performance between seen (89.94%) and OOD (88.43%) settings, with only a 1.51% degradation, demonstrating the robustness of evidence-guided reasoning under distribution shifts.

Zoom-Level Robustness. We evaluate MapDec-R1 across different tile zoom levels and map sources. As shown in Table 5, MapDec-R1 achieves over 96% average accuracy at zoom levels 14–16 and 94.64% overall accuracy across all scales, demonstrating robust detection under varying map resolutions. Performance decreases at zoom level 18, but the model remains effective across both seen and OOD sources.

Table 6: Ablation Study of the Two-Stage Evidence-Grounded Optimization Framework.

Model Variant	Stage 1 (Task)	Stage 2 (Evidence)	Accuracy	Macro-F1
w/o Stage 1	×	✓	0.7133	0.6992
w/o Stage 2	✓	×	0.8383	0.8358
MapDec-R1	✓	✓	**0.9200**	**0.9199**

5.3 Reasoning-Evidence Alignment Analysis

Global Evidence Alignment. We construct a Teacher Evidence Representation Space from atomic teacher evidence (Algorithm 1) and quantify reasoning–evidence discrepancy using SEMD. As shown in Fig. 4(a), MapDec-R1 reduces the distance from 0.539 to 0.453, a 16.0% reduction over the base VLM, indicating closer alignment with cartographic evidence patterns.

Dimension-wise Evidence Alignment. Using evidence prototypes defined at the appearance, element, and relational levels (Algorithm 2), we further measure dimension-wise cosine similarity. As shown in Fig. 4(b), MapDec-R1 achieves higher similarity across all three evidence dimensions, demonstrating consistent alignment with diverse cartographic evidence categories.

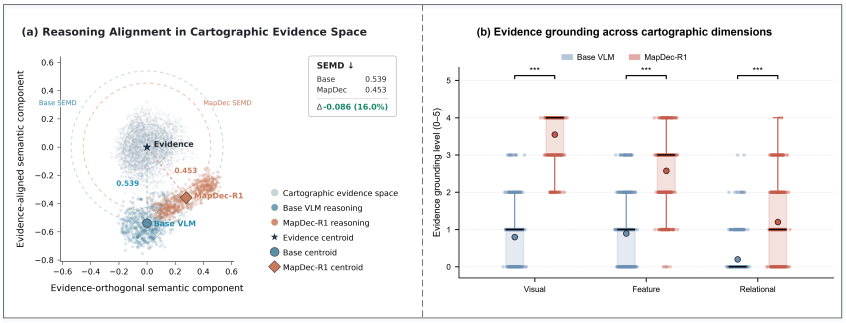

Figure 4: Evidence-grounded reasoning alignment analysis. (a) Reasoning distributions in the teacher-derived evidence representation space. (b) Alignment with Visual, Feature, and Relational evidence prototypes.

5.4 Ablation Studies

We investigate the contribution of each optimization stage by removing Stage 1 or Stage 2 individually. As shown in Table 6, MapDec-R1 achieves the best performance with both stages enabled. Removing Stage 1 reduces Accuracy and Macro-F1 by 20.67 and 22.07 percentage points, respectively, indicating that task-oriented initialization provides a necessary foundation for subsequent evidence-guided optimization. Removing Stage 2 leads to further drops of 8.17 and 8.41 percentage points, demonstrating the additional benefit of cartographic evidence-grounded refinement beyond outcome-level supervision.

6 Conclusion

This work proposes MapDec-R1, an evidence-grounded reasoning framework for AI-generated map detection. MapDec-R1 integrates structured cartographic evidence with evidence-grounded reasoning alignment and curriculum reinforcement learning to support reliable and interpretable authenticity assessment. Experimental results demonstrate that map authenticity can be effectively assessed through structured cartographic evidence involving semantic elements, representation rules, and spatial relationships, rather than relying solely on visual consistency. These findings validate our motivation that AI-generated map detection requires domain-specific evidence grounded in the intrinsic characteristics of maps, distinguishing it from general-purpose image detection. Future work will explore more scalable and adaptive evidence learning with reduced dependence on manually specified domain knowledge.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets and assisting manuscript writing, including language polishing, sentence refinement, and readability improvement. It is important to note that generative AI tools were not used for developing research ideas, formulating methodology, designing experiments, conducting scientific analysis, or interpreting experimental results. All research concepts, algorithmic design, experimental settings, and conclusions were developed, verified, and analyzed by the authors. The contributions of generative AI tools were limited to manuscript refinement and synthetic data generation assistance, with all AI-assisted outputs manually reviewed and validated by the authors. The authors take full responsibility for the final content of this manuscript, including all text, claims, and artifacts produced with the assistance of generative AI tools. We have ensured that all AI-assisted materials were carefully reviewed to maintain correctness, relevance, and compliance with ethical research practices.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No human subjects or animal experiments were involved in this study. All datasets used, including TMap-DD, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work. The MapDec-R1 framework, including the Cartographic Evidence Space, reward functions, and two-stage curriculum reinforcement learning strategy, is described in Section 4. Detailed training configurations, optimization settings, reward designs, and implementation details are provided in Section 5.1 and Appendix A.3. Baseline models and evaluation settings are summarized in Table 10. The construction procedures, generation configurations, and evaluation protocols of TMap-DD are documented in Section 3 and Appendix A.1. The teacher prompt templates used for evidence generation are provided in the supplementary materials. The supplementary materials include key source code, representative training dataset samples, and evaluation dataset samples to facilitate method verification and further research exploration.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Black Forest Labs. FLUX, 2024. URL https://github.com/black-forest-labs/flux.

Hongfei Cai, Chi Liu, Sheng Shen, Youyang Qu, and Peng Gui. Robust ai-synthesized image detection via multi-feature frequency-aware learning. In *Knowledge Science, Engineering and Management*, pp. 157–171. Springer, 2025.

Xu Chen, Songqiang Chen, Tian Xu, Bangguo Yin, Jian Peng, Xiaoming Mei, and Haifeng Li. Smapgan: Generative adversarial network-based semisupervised styled map tile generation method. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4388–4406, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Jie Gao, Marco Micheletto, Giulia Orrù, Sara Concas, Xiaoyi Feng, Gian Luca Marcialis, and Fabio Roli. Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. *Engineering Applications of Artificial Intelligence*, 133:108450, 2024.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

Google DeepMind. Gemini 3 flash preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview`, 2025. Google AI for Developers documentation.

Google DeepMind. Gemini 3.8 Flash Model Card, 2026a. URL `https://deepmind.google/models/model-cards/gemini-3-8-flash/`.

Google DeepMind. Nano Banana 2, 2026b. URL `https://gemini.google/overview/image-generation/`.

Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 814–825, 2025.

Yueqing Hu, Xinyang Peng, Shuting Peng, Hanqi Wang, and Tianhong Wang. Hán dān xué bù (mimicry) or qīng chū yú lán (mastery)? a cognitive perspective on reasoning distillation in large language models, 2026. URL `https://arxiv.org/abs/2601.05019`.

Tai-Ming Huang, Wei-Tung Lin, Kai-Lung Hua, Wen-Huang Cheng, Junichi Yamagishi, and Jun-Cheng Chen. Thinkfake: Reasoning in multimodal large language models for ai-generated image detection. *arXiv preprint arXiv:2509.19841*, 2025a.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28831–28841. IEEE, 2025b.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5967–5976. Ieee, 2017.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, et al. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27, 2025.

Yuhao Kang, Song Gao, and Robert E. Roth. Artificial intelligence studies in cartography: a review and synthesis of methods, applications, and ethics. *Cartography and Geographic Information Science*, 51(4):599–630, 2024.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Shuchang Lyu, Baoyuan Wu, and Guangliang Cheng. Raidx: A retrieval-augmented generation and grpo reinforcement learning framework for explainable deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11746–11755. ACM, 2025.

Yue Lin and Bo Zhao. Posthuman cartography? rethinking artificial intelligence, cartographic practices, and reflexivity. *Annals of the American Association of Geographers*, 115(3):499–512, 2025.

Hugo Touvron Llama. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2407.21783*, 2023.

Xiaochi Ma, Yifan Zhang, Wenbo Zhang, and Wenhao Yu. Geomapgan: Remote sensing image to map via two-stream network and adaptive noise. *Transactions in GIS*, 29(7):e70124, 2025.

Arpan Mahara, Liangdong Deng, and Naphtali D Rishe. Terrafly-forensics: A dataset for forensic detection of generated map images with quality assessment of generative models. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 558–563. IEEE, 2025.

Gengchen Mai, Ni Lao, Jielu Zhang, Lishen Mao, Zhangyu Wang, Nemin Wu, Krzysztof Janowicz, Koutian Wu, Jinmeng Rao, Song Gao, et al. On the ethics of generative geoai: Explainability, bias, hallucination, accountability, privacy, and trust. *Geography According to Foundation Models*, pp. 215–232, 2026.

Valentin Meo. Deepfake geography: a novel detection method for identifying manipulated satellite images. *Cartography and Geographic Information Science*, 53(4):465–478, 2026.

Microsoft Corporation. MAI-Image-2.5-Pro Model Card, 2026. URL https://microsoft.ai/pdf/MAI-Image-2.5_Pro-Model_Card.pdf.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

OpenAI. GPT-Image-1.5, 2025. URL https://openai.com/index/new-chatgpt-images-is-here/.

OpenAI. GPT-5.6 Sol, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol.

Jaqueline Pisetta, Fabíola A Souza, Jaqueline Amorim, Nathan D Antonio, Darlan M Nunes, Hideo Araki, and Silvana P Camboim. Mapping with words: Integrating large language models into geospatial practice. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:285–291, 2026.

Timothy J. Prestby. Trust in maps: What we know and what we need to know. *Cartography and Geographic Information Science*, 52(1):1–18, 2025a.

Timothy J. Prestby. Situating trust in cartography: Why do people trust maps and what does trust in maps mean? *Cartographic Perspectives*, (106):10–27, 2025b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Lei Tan, Shuwei Li, Mohan Kankanhalli, and Robby T Tan. Aggregating diverse cue experts for ai-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9314–9322, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b.

Fei Wu, Guanghao Ding, Zijian Niu, Zhenrui Wang, Lei Yang, Zhuosheng Zhang, and Shilin Wang. Generalizable detection of ai generated images with large models and fuzzy decision tree. *arXiv preprint arXiv:2603.28508*, 2026.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256. IEEE, 2023.

Sidi Wu, Yizi Chen, Samuel Mermet, Lorenz Hurni, Konrad Schindler, Nicolas Gonthier, and Loic Landrieu. Stegogan: Leveraging steganography for non-bijective image-to-image translation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7922–7931. IEEE, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

YuHao Xu, ZhenHai Wang, ZhiRu Wang, YunLong Guo, Rong Fan, HongYu Tian, and Xing Wang. Simdcl: dropout-based simple graph contrastive learning for recommendation. *Complex & Intelligent Systems*, 9(5):4751–4763, 2023.

Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. *ACM Transactions on Graphics (TOG)*, 44(4):1–19, 2025.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Kidus Zewde, Simiao Ren, Xingyu Shen, Jiaqi Wu, Yuchen Zhou, Tommy Duong, Zikang Zhang, Ethan Traister, and Kewen Xie. Gpt-image-2 in the wild: A twitter dataset of self-reported ai-generated images from the first week of deployment. *arXiv preprint arXiv:2604.25370*, 2026.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

Wenbo Zhang, Yifan Zhang, Hongyu Hao, Jianfeng Lin, Qingfeng Guan, and Wenhao Yu. Mapgenerator: a framework for learning a diffusion model for text promptable map generation. *Cartography and Geographic Information Science*, pp. 1–23, 2025a.

Yifan Zhang, Wenbo Zhang, Ziyi Zeng, Keying Jiang, Jingxuan Li, Wen Min, Wei Luo, Qingfeng Guan, Jianfeng Lin, and Wenhao Yu. Mapreader: a framework for learning a visual language model for map analysis. *International Journal of Geographical Information Science*, 39(9):1973–2008, 2025b.

Bo Zhao, Shaozeng Zhang, Chunxue Xu, Yifan Sun, and Chengbin Deng. Deep fake geography? when geospatial data encounter artificial intelligence. *Cartography and Geographic Information Science*, 48(4):338–352, 2021.

Chen Zhao, Wei-Ling Cai, and Zheng Yuan. Spectral normalization and dual contrastive regularization for image-to-image translation: C. zhao et al. *The Visual Computer*, 41(1):129–140, 2025.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18746–18758. IEEE, 2025.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 2242–2251. Ieee, 2017.

A APPENDIX

A.1 TMAP-DD-EVAL BENCHMARK DETAILS

TMap-DD-Eval is designed to evaluate the generalization capability of map authenticity detection models under diverse distribution shifts. This appendix provides detailed configurations of synthetic generation sources, real map scales, and geographic coverage. The synthetic generation configurations are summarized in Table 7, while the source and zoom-level distributions of real map tiles are presented in Table 8. The geographic coverage of real map samples is further detailed in Table 9.

Table 7: Synthetic map generation configurations in TMap-DD-Eval.

Paradigm	Generation family	Generator	Status	Samples
T2I	Commercial-based	Nano Banana 2 (Google DeepMind, 2026b)	Seen	60
		GPT-Image-2 (Zewde et al., 2026)	Unseen	37
		MAI-Image-2.5-Pro (Microsoft Corporation, 2026)	Unseen	25
		Doubao Seedream series (Seedream et al., 2025; Gao et al., 2025)	Unseen	8
	Diffusion-based	MapGenerator (Zhang et al., 2025a)	Seen	66
		Z-Image-Turbo (Z-Image Team et al., 2025)	Seen	30
		FLUX.1-dev (Black Forest Labs, 2024)	Seen	8
		SD 3.5 (Esser et al., 2024)	Unseen	16
I2I	GAN-based	CycleGAN (Zhu et al., 2017)	Seen	10
		GeoMapGAN (Ma et al., 2025)	Seen	5
		StyleGAN2-ADA (Karras et al., 2020)	Unseen	10
		Pix2Pix (Isola et al., 2017)	Unseen	9
		CAST (Yao et al., 2025)	Unseen	16

Table 8: Source and zoom-level distribution of real map tiles in TMap-DD-Eval.

Zoom level	Google Maps	OpenStreetMap	Total
z14	46	7	53
z15	44	20	64
z16	81	33	114
z17	26	7	33
z18	23	13	36
Total	**220**	**80**	**300**

Table 9: Geographic distribution of real map tiles in TMap-DD-Eval.

Region	Country	Source
Africa	Egypt	Google Maps, OpenStreetMap
Asia	Singapore	Google Maps
	South Korea	Google Maps
	Japan	Google Maps
Europe	United Kingdom	Google Maps, OpenStreetMap
	Netherlands	Google Maps, OpenStreetMap
	Italy	Google Maps, OpenStreetMap
	France	Google Maps, OpenStreetMap
North America	United States	Google Maps, OpenStreetMap

A.2 BASELINE MODELS

We evaluate representative open-weight and proprietary VLMs spanning different model scales, generations, and architecture families, as summarized in Table 10.

Table 10: Baseline models and comparison settings.

Model	Params.	Access	Adaptation	Baseline group
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	3B	Open-weight	Zero-shot	Backbone
Qwen2.5-VL-72B-Instruct (Bai et al., 2025b)	72B	Open-weight	Zero-shot	Scale-up
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	4B	Open-weight	Zero-shot	Next-generation
Qwen3.8-27B (Qwen Team, 2026)	27B	Open-weight	Zero-shot	Next-generation
InternVL3.5-8B-Instruct (Wang et al., 2025a)	8B	Open-weight	Zero-shot	Cross-family
Llama-3.2-11B-Vision-Instruct (Llama, 2023)	11B	Open-weight	Zero-shot	Cross-family
Gemma 3-27B-IT (Team et al., 2025)	27B	Open-weight	Zero-shot	Cross-family
GPT-5.6 Sol (OpenAI, 2026)	—	Proprietary	Zero-shot	Proprietary
Gemini 3.8 Flash (Google DeepMind, 2026a)	—	Proprietary	Zero-shot	Proprietary
DeepSeek V4.1 Flash (Xu et al., 2026)	—	Proprietary	Zero-shot	Proprietary
MapDec-R1	3B	Open-weight	Trained	Ours

A.3 Implementation Details

Additional optimization, sampling, and reward configurations for the two-stage training are summarized in Table 11. Stage 1 is initialized from Qwen2.5-VL-3B-Instruct, while Stage 2 continues from the merged Stage-1 checkpoint. The two stages share the same LoRA configuration and optimizer schedule, but differ in gradient accumulation, completion length, KL regularization, and reward design.

Table 11: Detailed configurations for the two-stage training.

Setting	Stage 1	Stage 2
Initialization	Qwen2.5-VL-3B-Instruct	Stage 1 merged checkpoint
Policy optimization	GSPO	GSPO
Training epochs	1	1
LoRA target	All linear layers	All linear layers
LoRA rank	8	8
LoRA alpha	32	32
LoRA dropout	0.05	0.05
Learning rate	2×10^{-5}	2×10^{-5}
Optimizer	AdamW	AdamW
Weight decay	0.1	0.1
LR scheduler	Cosine	Cosine
Warmup ratio	0.05	0.05
Per-device batch size	2	2
Gradient accumulation	8	4
Precision	BF16	BF16
Gradient checkpointing	✓	✓
Max. input length	4096	4096
Responses per prompt	8	8
Temperature	0.9	0.9
Max. completion length	256	512
KL coefficient β	0.01	0.02
Reward scaling	Group-based	GDPO-based
Reward functions	R^{acc}, R^{fmt}	R^{acc}, R^{fmt}, R^{ceg}, R^{len}
Reward weights	1.0; 0.1	1.0; 0.02; 0.7; 0.1
Accuracy scheduling	Fixed	R^{acc} ramped by $\gamma_{\text{acc}}(t)$

A.4 Evidence Alignment Evaluation

A.4.1 Teacher Evidence Space Alignment

Given reasoning representations and teacher evidence representations, we compute their centroid distance to measure distributional alignment.

Algorithm 1 Evidence Manifold Alignment Measurement

Require: Teacher evidence set $\mathcal{E}$, base reasoning set $\mathcal{R}_b$, MapDec-R1 reasoning set $\mathcal{R}_m$
Ensure: Evidence manifold distances D_b, D_m
Initialize evidence embedding set $\mathcal{Z}_e \leftarrow \emptyset$
for each evidence text $e \in \mathcal{E}$ **do**
    Encode e into embedding z_e
    $\mathcal{Z}_e \leftarrow \mathcal{Z}_e \cup \{z_e\}$
end for
Extract semantic prototypes: $\mathcal{P}_e \leftarrow \mathrm{PrototypeSelect}(\mathcal{Z}_e)$
Compute evidence centroid: $\mu_e \leftarrow \mathrm{Mean}(\mathcal{P}_e)$
for each reasoning set $\mathcal{R} \in \{\mathcal{R}_b, \mathcal{R}_m\}$ **do**
    Split reasoning into sentences $\mathcal{S}$
    Encode sentences into embeddings $\mathcal{Z}_s$
    Select top-k evidence-related sentences: $\mathcal{S}^* \leftarrow \mathrm{TopK}(\mathrm{CosSim}(\mathcal{Z}_s, \mu_e))$
    Aggregate selected representations: $z_r \leftarrow \mathrm{MeanPool}(\mathcal{Z}_s^*)$
    Compute reasoning centroid μ_r
    Calculate SEMD: $D \leftarrow \|\mu_r - \mu_e\|_2$
end for
Assign: $D_b \leftarrow D(\mathcal{R}_b)$, $D_m \leftarrow D(\mathcal{R}_m)$
return D_b, D_m

A.4.2 Evidence Prototype Alignment

Given evidence prototypes defined in the Cartographic Evidence Space, we compute the cosine similarity between reasoning representations and each evidence dimension.

Algorithm 2 Cartographic Evidence Attribution Evaluation

Require: Student reasoning R, evidence dimensions $\mathcal{D} = \{\mathrm{Visual}, \mathrm{Feature}, \mathrm{Relational}\}$
Ensure: Evidence scores $\{s_d \mid d \in \mathcal{D}\}$
Construct evaluation prompt according to evidence dimension definitions
for each dimension $d \in \mathcal{D}$ **do**
    Evaluate whether R contains concrete cartographic observations for d
    Assign evidence score: $s_d \leftarrow \mathrm{Judge}(R, d)$
end for
Validate scores within range $[0, 5]$
return $\{s_d \mid d \in \mathcal{D}\}$

A.5 Category-Aware Teacher Prompt Templates

We employ category-aware prompts to generate teacher cartographic reasoning references. Different prompts are designed for real and AI-generated 2D map tile images, respectively, as shown in Fig. 5 and Fig. 6. The prompts guide the teacher model to discover authenticity-related evidence according to the Cartographic Evidence Space.

The prompt defines the map authenticity verification task and structures the analysis around the three dimensions of the Cartographic Evidence Space, with category-level reasoning chains and output constraints designed to ensure observable, evidence-grounded, and reusable reasoning.

Prompts for ai-generated vector map tiles

Task Definition
You are an expert in AI map authenticity analysis. You will be provided with a vector map tile that is entirely generated by AI.
Your task is to construct a complete reasoning process from "image observation" to "forgery judgment," and finally produce an interpretable and reusable inference result that can serve as a summarized output for model learning.
Basic Constraints (Must Be Strictly Followed)
- Only rely on phenomena directly observable in the image
- No unsupported speculation or subjective assumptions
- If no anomaly is found, explicitly state "no anomaly observed"
- Avoid vague descriptions; all statements must refer to specific visual evidence in the image

Cartographic Evidence Space

Technical Analysis and Map Perception (Sources of Evidence)
During analysis, you must actively examine the image from multiple perspectives, rather than focusing only on the most obvious anomalies.
You may refer to the following aspects to identify potential anomalies (record only "observable phenomena," do not draw conclusions at this stage):

Visual Appearance Level

I. Rendering Style and Visual Consistency
- Inconsistent styles within the same feature class: Within the same tile, features of the same type (roads, buildings) showing mixed styles (e.g., solid/dashed lines mixed, inconsistent fill patterns).
- Color anomalies: Local abrupt shifts in color, saturation, or brightness; overall overly saturated or dull; color assignments violating map conventions (e.g., water rendered in red).
- Unreasonable shadows: Standard vector maps typically have no shadows. If shadows are present, assess whether they are consistent with terrain and light source.
- Cultural confusion in symbol systems: Mixing cartographic conventions from different regions (e.g., simultaneous left-hand and right-hand traffic arrows), symbols appearing deliberately stitched together.
- Map style patching: Abrupt switching of basemap style, color scheme, or rendering resolution within a local area, creating a visible collage-like boundary.

II. Fill and Edge Artifacts
- Texture repetition / tiling artifacts: Region fills (e.g., green spaces, water) showing overly perfect periodic patterns, lacking the subtle randomness of real rendered maps (i.e., over-regularization).
- Edge artifacts: Edges exhibiting unnatural jaggies (excessive stair-stepping), over-sharpening or blurring, ghosting/double-image effects (e.g., double shadow along road lines or text edges).
- Abnormal smoothing: Vector-rendered edges should be sharp. Excessive smoothness may be caused by AI interpolation.
- Unnatural background: Tile edges or corners showing repeating patterns or blur, foreground and background blending unnaturally.

III. Text Labels and Symbol System Anomalies
- Map symbol misuse: Functional symbols do not match the actual regional semantics, e.g., hospital symbols placed in factory areas, industrial pollution symbols near schools, highway numbering styles applied to ordinary county roads, or different road classes using identical numbering systems and styles. Pay attention to whether the type and location of any icon/symbol are logically consistent with its surroundings.
- Text content issues: Garbled characters, missing strokes, unreadable symbols; or "text-like" pseudo-text textures; presence of non-existent place names, anomalous word combinations, or random character concatenations.
- Anomalous multilingual mixing: Illogical mixing of Latin characters, fictitious foreign languages, or random concatenations from different language systems within a predominantly Chinese map area.
- Text deformation and layout: Text twisted, distorted, or not aligned along road curves (for road names).
- Inconsistent styling: Inconsistent font, size, color, or spacing for labels of the same category.
- Poor text-background integration: Insufficient color contrast, or text strangely truncated or occluded by other features.

Cartographic Element Level

IV. Road Network and Geometric Topology Anomalies
- Road connectivity topology issues: Unreasonable dead-end roads (no physical reason to end), abnormal intersection angles (overly acute, random irregular angles), self-intersecting roads, dangling lines, duplicate vertices, irregular or overly regular spacing between parallel roads.
- Unnatural intersection of linear features: Unreasonable crossings or overlaps of linear features such as railways, rivers, and highways on the same planar level, lacking necessary engineering structures such as overpasses or tunnels.
- Unreasonable connections due to data inconsistency: Dead-ends or connection anomalies caused by mixing old and new road data (e.g., newly built roads connecting to abandoned ones), inconsistent update traces across regions.
- Building footprint distortion: Irregularly twisted contours, self-intersection, non-closure, bizarre non-rectangular shapes, severely disproportionate scale relative to surroundings.
- Road curve smoothness issues: Road curves appearing polyline-like, jagged, rather than smooth splines.
- Natural feature morphology anomalies: Natural curves such as rivers and coastlines lack realistic random perturbations and fractal characteristics, appearing overly smooth, geometric, rigid, or showing tendencies toward random noise patterns.
- Local geometric anomalies: Unnatural twisting, jittering, or other micro-scale irregularities at road edges, building corners, and boundary lines.

V. Noise, Frequency, and Statistical Regularity Anomalies
- Abnormal noise patterns: Presence of regular noise dot matrices globally or locally, or complete absence of noise (overly clean appearance).
- Abnormal spatial statistical distribution: Spatial structures such as building arrangement, road spacing, and green space distribution exhibit over-regularized tendencies, lacking the randomness, hierarchy, and natural variation of real urban spaces.
- Localized generative repetition: Highly similar or even mirrored building clusters or block layouts appearing in different areas, reflecting the template reuse characteristic of generative models.

Relational Level

VI. Spatial Relationships and Semantic Logic Errors
- Implausible spatial relationships: Roads passing directly through buildings, rivers and roads overlapping without bridge symbols, roads abruptly cut off at tile edges without reasonable continuation.
- Isolated entities and spatial disconnection: Buildings, vegetation patches, or other elements isolated in the middle of lakes, within road right-of-way, or in other implausible spatial locations, lacking normal spatial connectivity with the surroundings; buildings not grounded to the surface, elements harshly disconnected from surroundings, isolated and floating without connection.
- Semantic and land-use incoherence: Absurd semantic relationships (e.g., road labels on water bodies, industrial symbols near schools, docks in inland areas, forest over deep water).
- Scale and spatial proportion anomalies: Noticeable distortion in the size ratio among buildings, roads, green spaces, etc. within the same tile, e.g., a basketball court area abnormally larger than a large building, or secondary roads appearing wider than arterial roads.

Reasoning Chain
You must conduct an exhaustive class-by-class scan of the image based on the six anomaly categories above. For each category, proactively search the image for any observable anomaly. Do not stop after discovering only a few obvious issues; your goal is to identify all visually detectable anomalies in the image.
Reasoning Chain Format (Must Follow)
Organize reasoning chains by category: at most one reasoning chain per category, summarizing all specific anomalies observed within that category. If no anomaly is found in a category, explicitly state "No anomaly observed in this category."
Reasoning Chain N
- **Observation**: Describe objectively observable phenomena
- **Anomaly**: Identify the type of anomaly
- **Validation**: Rule out reasonable explanations (e.g., design choices or resolution issues)
- **Evidence**: Confirm it as evidence of AI generation
If uncertain → mark as "uncertain" and do not treat it as evidence
Deep Inspection
After completing the initial analysis, further consider:
- Are there subtle geometric anomalies (jitter, distortion)?
- Are there local repetitions or excessive regularity?
- Are there slight inconsistencies in style?
If new issues are found, add additional reasoning chains.
Final Response Organization
Based on the above reasoning, organize the content into a coherent and complete response.
Requirements:
Output in JSON format, including the following fields:
- thoughts: A detailed description of the reasoning process that follows the logic of "Observation → Anomaly → Validation → Evidence", and is structured in an "overall → detailed points → overall" format. The opening sentence should provide an overall judgment indicating that the image exhibits clear signs of AI generation. In the main body, use structured numbering such as "1. 2. 3. ..." to summarize and organize the anomaly characteristics across the entire reasoning chain. When describing each anomaly, specific visual details must be preserved, and generic descriptions are not allowed. The final sentence should provide a concluding statement that clearly determines the image to be AI-generated.
- answer: final conclusion (real/fake)
Final Output Format
Must follow the structure below:
- Reasoning Process (List all reasoning chains)
- Final Output (Provide the structured response including thoughts and answer fields)
Prohibited Actions
- Drawing conclusions without observational evidence
- Skipping reasoning steps
- Using vague or speculative language
- Forcing non-existent issues
- Fabricating information

Figure 5: Teacher prompt template for AI-generated vector map tiles.

Prompts for real vector map tiles

Task Definition
You are an AI map authenticity analysis expert. You will be provided with a vector map tile originating from a real source.
Your task is to analyze the image content, complete an authenticity verification, and generate an explainable and reusable reasoning result that can be used as a summarized response for model learning.
Analysis Procedure Constraints (Must Be Strictly Followed)
You must:
Cartographic Evidence Space
1. Examine the following 6 categories in order (do not skip any):
Visual Appearance Level
1) Rendering Style & Visual Consistency
- Style uniformity: Similar features (roads, buildings, water bodies) share consistent styles (color, line type, fill style).
- Color plausibility: Colors align with map conventions (e.g., water is blue, vegetation is green), with no abrupt changes.
- No abnormal shadows: Standard vector maps typically have no shadows, or shadows are consistent.
- Symbol system consistency: Symbols follow a unified standard, with no mixing of styles from different mapping systems.
2) Fill & Edge Quality
- Natural fill: Area fills (vegetation, water) appear natural, without obvious repetition or periodic patterns.
- Edge clarity: Edges are sharp, without jaggedness, blurring, or ghosting.
- No interpolation artifacts: No abnormal smoothing or AI-generated transitional blurring.
- Background consistency: The overall background is clean, without stitching or repetition artifacts.
3) Noise & Image Quality
- Noise plausibility: The image is clean but not overly "perfect," consistent with vector rendering characteristics.
- No abnormal frequency patterns: No periodic noise or unusual textures.
Cartographic Element Level
4) Road Network & Geometric Topology
- Road connectivity rationality: Roads connect naturally, with no meaningless dead ends; intersection angles follow engineering norms (approximately right angles or smooth transitions); no self-intersections or dangling lines.
- Topological integrity: The road network has good connectivity, with no obvious topological errors (e.g., duplicate nodes, abnormal overlaps).
- Building geometry regularity: Building outlines are closed, structurally regular, and conform to common shapes (rectangles, polygons, etc.).
- Curve smoothness: Road curves are smooth and natural, without noticeable jagged or polyline artifacts.
- Local geometric stability: Details (edges, corners) show no jitter or distortion.
5) Text Labels
- Text clarity and readability: No garbled or missing characters.
- Reasonable placement: Text is placed logically along roads or within regions.
- Style consistency: Fonts, sizes, and colors are consistent for similar labels.
- Proper contrast: Text contrasts clearly with the background, without abnormal occlusion or blending.
Relational Level
6) Spatial Relationships & Semantic Logic
- Spatial consistency: Roads do not pass through buildings (unless justified by structures like bridges); relationships between rivers and roads are logical.
- Natural feature relationships: Buildings, roads, and water systems interact in a way consistent with real-world logic.
- No isolated elements: No floating or unconnected features.
- Semantic consistency: Labels correspond correctly to features (e.g., road names align with roads).
2. Each category must include at least one reasoning chain, but it may be marked as "Not Observable."
Reasoning Chain Format (Strict Template)
Case 1: Observable
Reasoning Chain N (Category Name)
- Observation: Describe specific visible phenomena in the image (must be observable)
- Consistency: Explain how it conforms to real map characteristics
- Validation: Explain why AI has difficulty generating this feature consistently
- Evidence: Confirm it as evidence of a real image
Case 2: Not Observable (Must Be Used)
Reasoning Chain N (Category Name)
- Observation: The feature of this category cannot be observed in the current image (state the reason, e.g., resolution/content limitation)
- Consistency: Not applicable
- Validation: Cannot be verified
- Evidence: Uncertain (not used as evidence)
Key Constraints
- Do not speculate about unseen content
- Do not infer or complete missing information
- Do not use phrases like "possibly exists" or "should have"
- All observations must correspond to specific visible regions
Deep Inspection
Additionally check for:
- Minor irregularities present in real data
- Imperfections (e.g., slight offsets, asymmetry)
Final Response Structuring
Based on the complete reasoning process above, reorganize the content into a coherent, complete, and specific response.
Requirements:
Output in JSON format with the following fields:
- **thoughts**: Provide a detailed description of the reasoning chains, organized in a **general–specific–general** structure. The opening sentence should present an overall judgment, summarizing that the image exhibits multiple authenticity characteristics. The middle section should be developed in a numbered format (1. 2. 3. ...), reorganizing and synthesizing the existing reasoning chains. It should focus on explicitly describing the visual observational details and explaining how these details reflect the characteristics of real vector map tiles (without requiring one-to-one correspondence with each reasoning chain). Generalized summaries or abstract conclusions are not allowed. The conclusion should provide a clear final judgment, confirming that the image is a real map tile and not AI-generated.
- **answer**: Final conclusion (real/fake)
Final Output Format
Must follow this structure:
Reasoning Process(List all reasoning chains)
Final Output(Provide the structured response containing thoughts and answer fields)

Figure 6: Teacher prompt template for real vector map tiles.