WHEN DOES EMBEDDING ARITHMETIC FAIL? A SYSTEMATIC ANALYSIS IN REMOTE SENSING VISION-LANGUAGE MODELS

Jinpyo Hong[1] **Le Yu**[2]
[1]School of Engineering, Brown University, Providence, RI, USA
[2]Ministry of Education Key Laboratory for Earth System Modeling,
Department of Earth System Science, Tsinghua University, Beijing 100084, China

ABSTRACT

Embedding arithmetic promises flexible compositional queries over remote sensing imagery—transforming a harbor into an airport by subtracting "water" and adding "runway"—yet when this actually works remains poorly understood. We systematically evaluate four CLIP-based models across five RS datasets and identify concept entanglement as the dominant failure mode (40–60% of failures): semantically related concepts occupy overlapping embedding subspaces that confound arithmetic. We propose a pre-hoc entanglement metric—requiring only text embeddings—that predicts failure with AUC up to 0.818, with GeoRSCLIP showing the most consistent predictions (mean AUC=0.675). Notably, embedding geometry does not reliably predict compositional capability (r=0.30, p=0.20), suggesting discriminative and compositional reasoning require different representational properties. We provide practical guidelines: arithmetic succeeds for well-separated concepts (88%) but fails predictably for structurally similar classes (42%).

1 INTRODUCTION

Practitioners increasingly want to use embedding arithmetic for compositional queries over remote sensing imagery—computing transformations like harbor $-$ water $+$ runway $\approx$ airport using precomputed embeddings from foundation models. This capability, inspired by word2vec analogies (Mikolov et al., 2013) and extended to vision-language models (Radford et al., 2021), promises flexible scene reasoning without additional training.

Vision-language models (VLMs) have transformed remote sensing (RS) through zero-shot classification and open-vocabulary detection (Huo et al., 2025; Zhang et al., 2024). Models like RemoteCLIP (Liu et al., 2024) and GeoRSCLIP provide precomputed embeddings for diverse downstream tasks. Yet embedding arithmetic's effectiveness in RS remains poorly characterized.

RS imagery presents unique challenges: domain-specific semantics, fine-grained distinctions, and structural similarities that confound arithmetic (Figure 1). We provide the first comprehensive analysis of when and why embedding arithmetic fails across multiple RS datasets and VLMs.

Our contributions: (1) evaluation framework spanning 5 datasets, 4 models, and 27–34 arithmetic experiments per configuration under both centroid-based and per-image retrieval; (2) failure taxonomy identifying concept entanglement as dominant (40–60%); (3) pre-hoc entanglement metric for failure prediction using only text embeddings; (4) exploratory 2×2 capacity vs. domain analysis; and (5) practical guidelines.

2 METHODOLOGY

2.1 EMBEDDING ARITHMETIC FRAMEWORK

Given a reference image embedding $\mathbf{e}_{\text{ref}}$ from class C_{ref}, we perform arithmetic:

$$\mathbf{e}_{\text{result}} = \text{normalize}\left(\mathbf{e}_{\text{ref}} - \alpha \cdot \mathbf{t}_{\text{sub}} + \alpha \cdot \mathbf{t}_{\text{add}}\right) \tag{1}$$

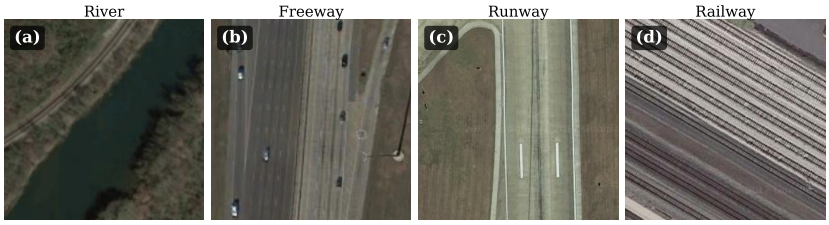

Figure 1: Structural confusion in Remote Sensing: river, freeway, runway, and railway share linear morphology despite distinct semantics, creating challenges for embedding arithmetic.

where $\mathbf{t}_{\text{sub}}$ and $\mathbf{t}_{\text{add}}$ are text embeddings of concepts to subtract/add, and $\alpha = 3.0$ controls modification strength (see Appendix E for sensitivity analysis). Success is determined by nearest-neighbor retrieval among class centroids (see Appendix C for formal definition). We use FAISS (Johnson et al., 2019) for efficient retrieval. We evaluate under two complementary settings. **Centroid-based retrieval** uses class-averaged embeddings as both reference inputs and retrieval targets, providing a stable baseline that isolates concept manipulation effects from instance-level variation. **Per-image retrieval** samples $N = 50$ images per experiment, performs arithmetic on each, and retrieves against all dataset images with majority voting ($k = 10$), capturing instance-level variance that reflects real-world scenarios. Bootstrap confidence intervals (95%, 1,000 samples) quantify uncertainty.

2.2 Failure Taxonomy

Building on insights from compositional reasoning failures (Yuksekgonul et al., 2022; Thrush et al., 2022), we categorize failures into four mutually exclusive types using a **post-hoc classification criterion**. This criterion examines the relationship between the *predicted* class and the concepts involved *after* an experiment has been run: **Entanglement failure**: predicted class is semantically related to reference or target (cosine similarity >0.5). **No-effect failure**: prediction equals reference. **Structural confusion**: predictions share visual patterns but differ semantically. **Other failures**: predictions unrelated to reference, target, or concepts.

Note: This post-hoc taxonomy differs from the **pre-hoc entanglement metric** (Section 3.3), which predicts failures *before* running experiments.

3 Experiments

3.1 Datasets and Models

We evaluate on five standard RS benchmarks: UCM (Yang & Newsam, 2010) (21 classes, 2,100 images), AID (Xia et al., 2017) (30 classes, 10,000 images), PatternNet (Zhou et al., 2018) (38 classes, 30,400 images), RSICD (Lu et al., 2017) (30 classes, 10,921 images), and EuroSAT (Helber et al., 2019) (10 land-cover classes, 27,000 Sentinel-2 images at 10 m resolution). We compare four CLIP-based models forming a 2×2 matrix to isolate capacity from domain effects: CLIP ViT-B/32 serves as our general-purpose baseline, RemoteCLIP ViT-B/32 represents RS-specific adaptation at smaller scale, CLIP ViT-L/14 provides larger capacity while remaining general-purpose, and GeoRSCLIP ViT-L/14 combines larger architecture with RS-specific pretraining on RS5M (5M images). All models are accessed via OpenCLIP (Ilharco et al., 2021).

Probe selection criteria. We design 27–34 arithmetic probes per dataset following three requirements: (1) reference and target classes must exist in the dataset's taxonomy, (2) the transformation must be semantically meaningful for RS scenes (e.g., land-use change, infrastructure modification), and (3) probes span difficulty levels approximately uniformly (7–10 probes per level). This systematic selection ensures coverage across concept types.

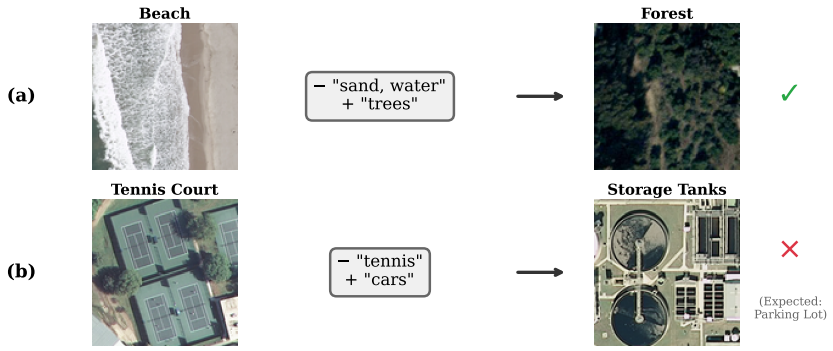

Figure 2: Embedding arithmetic examples. (a) Well-separated concepts (sand/water vs. trees) enable correct retrieval. (b) Entangled concepts (rectangular structures) cause retrieval of Storage Tanks instead of Parking Lot.

3.2 Main Results

Table 1 presents centroid-based success rates across all 20 model-dataset configurations. The results reveal that embedding arithmetic achieves moderate overall success (66–74% mean across models) but with substantial variation across datasets and model choices.

Table 1: Embedding arithmetic success rates (%) using centroid retrieval.

Model	UCM	AID	PatternNet	RSICD	EuroSAT	Mean
CLIP ViT-B/32	74.1	64.7	64.7	73.5	60.7	67.5
RemoteCLIP	74.1	76.5	50.0	**82.4**	46.4	65.9
CLIP ViT-L/14	74.1	**79.4**	79.4	70.6	60.7	72.8
GeoRSCLIP	**77.8**	73.5	**82.4**	67.6	**67.9**	**73.8**

Our exploratory 2×2 capacity-domain analysis (Appendix B) finds that the capacity effect (+6.6%) exceeds the domain effect (-0.3%), but neither reaches statistical significance ($p > 0.4$), precluding definitive conclusions. This suggestive evidence warrants further investigation with additional model pairs.

Despite this statistical uncertainty, one pattern is consistent: L/14 models outperform B/32 models in 4 of 5 datasets regardless of domain pretraining. No single model dominates across all datasets, indicating that optimal model selection depends on target dataset characteristics. GeoRSCLIP excels on fine-grained PatternNet (82.4%) and coarse-grained UCM (77.8%), while RemoteCLIP leads on caption-based RSICD (82.4%).

Table 2: Per-image retrieval success rates (%) with $N = 50$ images per experiment. The $\pm$ values report SD *across 27–34 probes*, reflecting difficulty diversity across transformations.

Model	UCM	AID	PatternNet	RSICD	EuroSAT	Mean
CLIP ViT-B/32	65.5±40.2	61.3±41.5	58.9±43.4	62.1±39.7	51.0±38.4	59.8
RemoteCLIP	70.9±38.5	75.3±35.9	46.2±43.3	78.7±33.0	43.4±43.8	62.9
CLIP ViT-L/14	66.8±39.3	71.5±34.8	78.8±36.9	68.5±37.8	63.4±43.4	69.8
GeoRSCLIP	74.1±37.9	71.2±40.4	79.1±35.6	69.7±40.7	**67.4**±40.2	**72.3**

Per-image retrieval (Table 2) better reflects real-world deployment, with success rates dropping 2–11% versus centroid retrieval. The reported SDs (35–43%) reflect *across-probe* difficulty diversity, not within-experiment noise; within-probe SD averages 14–20%, confirming systematic instance-level effects where prototypical images succeed more reliably than boundary outliers.

Fine-grained probes. Notably, PatternNet's 38 classes already include fine-grained probes that go beyond standard GIS/remote sensing categories. These reveal sharp performance contrasts: *forest→christmas_tree_farm* achieves only 2% success (vegetation subtype), while *residential→nursing_home* reaches 70% (built environment subtype), and *trans-*

former_station→solar_panel and *oil_well→storage_tank* both achieve 100% (energy/industrial subtypes). The key insight is that arithmetic fails most on fine-grained *within-domain* distinctions where concept entanglement is strongest.

EuroSAT reinforces this in the ecological domain: fine-grained vegetation probes (*AnnualCrop↔PermanentCrop*: 8–34%; *Forest→HerbaceousVegetation*: 0–4%) fail while the structural control *Highway→River* achieves 86–100%, confirming that entanglement—not task difficulty—drives failure.

3.3 Failure Analysis and Practical Guidelines

Across all configurations, concept entanglement emerges as the dominant failure mode (40–60% of failures), consistent with the challenges of learning disentangled representations (Locatello et al., 2019; Higgins et al., 2017): semantically related concepts share overlapping embedding subspaces that confound arithmetic (Figure 2).

Pre-hoc entanglement metric. To predict failures *before* running experiments, we evaluate a lightweight metric: $\text{sim}(\mathbf{t}_{\text{add}}, \mathbf{t}_{\text{sub}})$, the cosine similarity between add and subtract concept embeddings. This differs from the post-hoc taxonomy criterion, as pre-hoc metric requires only text embeddings and can be computed before any experiment.

Table 3: Pre-hoc failure prediction using $\text{sim}(\mathbf{t}_{\text{add}}, \mathbf{t}_{\text{sub}})$. Target AUC > 0.60 shown with †.

Model	UCM	AID	PatternNet	RSICD	EuroSAT	Mean
CLIP ViT-B/32	0.679†	0.650†	0.646†	0.314	0.625†	0.583
RemoteCLIP	0.673†	0.500	0.650†	0.127	0.800†	0.550
CLIP ViT-L/14	0.577	0.818†	0.368	0.533	0.700†	0.599
GeoRSCLIP	0.554	0.800†	0.525	0.727†	0.771†	**0.675**†
Overall mean across all 20 configurations:						0.602†

Table 3 shows mean AUC=0.602 across all configurations (range: 0.127–0.818), with twelve of twenty exceeding 0.60. GeoRSCLIP achieves the most consistent predictions (mean AUC=0.675), suggesting large-scale RS pretraining creates more separable concept structures.

Strikingly, embedding geometry shows weak positive correlation ($r = 0.30$, $p = 0.20$) with arithmetic success (Appendix C): RemoteCLIP achieves the highest centroid separation (0.646) yet not the highest arithmetic performance, suggesting compositional reasoning requires different representational properties than discriminative classification.

Practical guidelines.

Table 4 provides actionable thresholds based on our pre-hoc entanglement metric. Given variability across configurations, these are heuristics rather than guarantees.

Table 4: When to use embedding arithmetic based on pre-hoc entanglement score.

Confidence	Entanglement	Examples (Observed Success Rate)
High	< 0.5	forest→agricultural (88%)
Medium	0.5–0.7	harbor→airport (71%)
Low	> 0.7	river↔freeway (42%)

4 Conclusion

Concept entanglement is the dominant bottleneck for embedding arithmetic in RS VLMs (40–60% of failures). Our text-only pre-hoc metric predicts these failures (AUC up to 0.818), while embedding geometry—despite capturing discriminative quality—does not reliably predict compositional capability (r=0.30, p=0.20). **Limitations and future work.** Our capacity-domain analysis has confounds and does not reach significance; evaluation is transductive, closed-set, and uses manually curated probes. Extensions include multi-label evaluation (Sumbul et al., 2019) and location-aware arithmetic via geospatial embeddings (Klemmer et al., 2025; Vivanco Cepeda et al., 2023).

REFERENCES

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15338–15347, 2023.

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

Karim El Khoury, Maxime Zanella, Benoît Gérin, Tiffanie Godelaine, Benoît Macq, Saïd Mahmoudi, Christophe De Vleeschouwer, and Ismail Ben Ayed. Enhancing remote sensing vision-language models for zero-shot scene classification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13225–13234, 2024.

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

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Chunlei Huo, Keming Chen, Shuaihao Zhang, Zeyu Wang, Heyu Yan, Jing Shen, Yuyang Hong, Geqi Qi, Hongmei Fang, and Zihan Wang. When remote sensing meets foundation model: A survey and beyond. *remote sensing*, 17(2):179, 2025.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547, 2019.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4347–4355, 2025.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 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, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

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

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19305–19314, 2023.

Dylan Sam, Devin Willmott, Joao D Semedo, and J Zico Kolter. Finetuning clip to reason about pairwise differences. *arXiv preprint arXiv:2409.09721*, 2024.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *arXiv preprint arXiv:1902.06148*, 2019.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018.

A Related Work

Vision-Language Models for RS. Foundation models for RS have accelerated rapidly. RemoteCLIP (Liu et al., 2024) adapts CLIP through RS image-text pretraining, while GeoRSCLIP leverages RS5M (5M images) with ViT-L/14 architecture. Self-supervised approaches include SatMAE (Cong et al., 2022) for temporal/multi-spectral imagery and RingMo (Sun et al., 2022) as a general-purpose RS foundation model. These models excel at zero-shot classification but their compositional reasoning capabilities remain understudied.

Embedding Arithmetic. The theoretical foundations trace to Mikolov et al. (2013), who demonstrated word embeddings support analogical reasoning. Pennington et al. (2014) provided theoretical grounding through GloVe. Radford et al. (2021) extended these properties to multimodal embeddings. However, conditions under which arithmetic succeeds or fails remain poorly characterized for specialized domains.

Compositional Reasoning in VLMs. Yuksekgonul et al. (2022) showed VLMs often behave like bags-of-words, struggling with attribute binding. Thrush et al. (2022) introduced Winoground for probing compositionality. The challenge of disentangled representations has been explored by Higgins et al. (2017) and critically examined by Locatello et al. (2019).

Modality Gap in VLMs. Liang et al. (2022) showed VLMs exhibit a "modality gap" where image and text embeddings occupy distinct regions. El Khoury et al. (2025) explored bridging this gap for RS applications.

Composed Image Retrieval. Baldrati et al. (2023) introduced zero-shot composed retrieval using textual inversion. Saito et al. (2023) proposed Pic2Word for mapping images to pseudo-word tokens. Gu et al. (2024) developed LinCIR using language-only training. Sam et al. (2024) showed pairwise difference supervision improves analogical reasoning.

B Capacity vs. Domain Analysis

Table 5: 2×2 capacity vs domain analysis. Neither effect reaches significance ($p > 0.4$).

Capacity	Domain	Arithmetic (%)	Geometry
ViT-B/32	General	67.5	0.467
ViT-B/32	RS-Specific	65.9	0.646
ViT-L/14	General	72.8	0.443
ViT-L/14	RS-Specific	73.8	0.590
Capacity Effect (L/14 - B/32)		+6.6%	−0.04
Domain Effect (RS - General)		−0.3%	+0.16

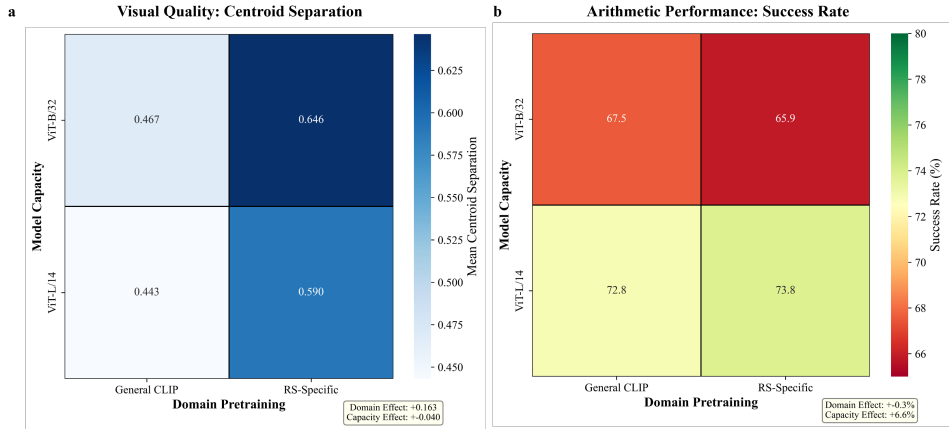

Figure 3: 2×2 capacity vs domain analysis. (A) Geometry: RS models achieve 36% higher centroid separation. (B) Arithmetic: Capacity effect (+6.6%) dominates domain effect (−0.3%).

C Geometry Metrics

Centroid computation. Class centroids are computed as:

$$\mathbf{p}_c = \frac{1}{|I_c|} \sum_{i \in I_c} \mathbf{e}_i \tag{2}$$

where I_c is the set of images belonging to class c.

Centroid separation measures inter-class discriminability:

$$\text{Sep} = \frac{2}{K(K-1)} \sum_{i<j} ||\mu_i - \mu_j||_2 \tag{3}$$

Uniformity measures distribution evenness on the hypersphere:

$$\text{Unif} = \log \mathbb{E}_{(i,j)\sim\text{pairs}} \left[e^{-2||\mathbf{e}_i - \mathbf{e}_j||_2^2} \right] \tag{4}$$

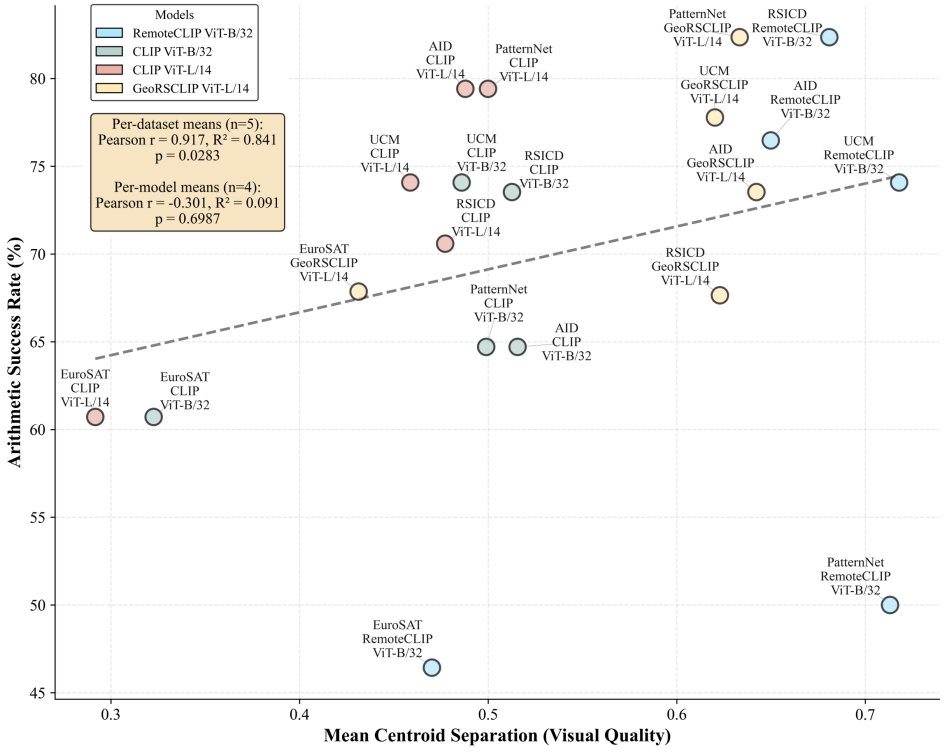

Figure 4: Geometry (centroid separation) vs arithmetic success shows weak correlation (r=0.30, p=0.20). Visual representation quality does not reliably predict compositional capability.

D Prompt Sensitivity Analysis

We evaluate with six prompt templates: "{concept}", "an image with {concept}", "a scene with {concept}", "the presence of {concept}", "an image containing {concept}", and "{concept} visible in the scene".

Table 6: Prompt sensitivity: mean variance (%) across 6 templates. Lower = more robust.

Model	UCM	AID	PatternNet	RSICD	Mean
CLIP ViT-B/32	7.2	5.6	2.7	0.0	3.9
RemoteCLIP	6.4	0.0	9.7	8.3	6.1
CLIP ViT-L/14	22.3	15.5	4.9	9.9	13.1
GeoRSCLIP	5.9	4.3	2.7	5.4	**4.6**

GeoRSCLIP achieves 2.9× better prompt robustness than CLIP L/14 (4.6% vs 13.1%), suggesting domain-specific pretraining creates more stable concept representations.

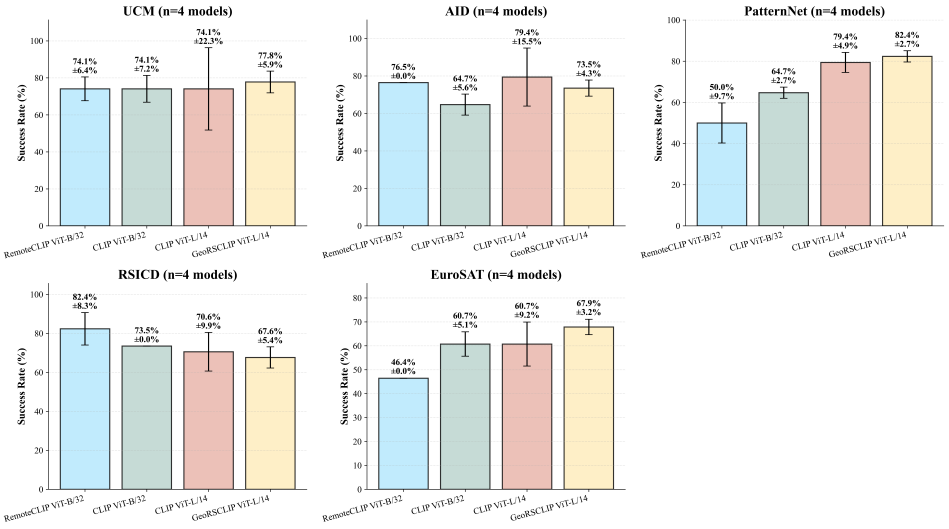

Figure 5: Arithmetic success rates with prompt sensitivity error bars (± 1 std across 6 templates).

E Alpha Sensitivity Analysis

Table 7: Alpha sensitivity. Performance plateaus at α=2.5–3.0; we use α=3.0.

Dataset	**Model**	α=0.5	α=1.5	α=2.5	α=3.0	α=5.0
UCM	CLIP B/32	7.4	59.3	74.1	74.1	74.1
UCM	RemoteCLIP	3.7	44.4	74.1	74.1	77.8
UCM	CLIP L/14	11.1	44.4	70.4	74.1	74.1
AID	CLIP B/32	0.0	47.1	64.7	64.7	67.6
AID	RemoteCLIP	0.0	47.1	79.4	76.5	79.4
AID	CLIP L/14	0.0	58.8	79.4	79.4	82.4
PatternNet	CLIP B/32	0.0	41.2	61.8	64.7	67.6
PatternNet	RemoteCLIP	0.0	32.4	41.2	50.0	50.0
PatternNet	CLIP L/14	0.0	67.6	82.4	79.4	88.2
RSICD	CLIP B/32	0.0	52.9	70.6	73.5	73.5
RSICD	RemoteCLIP	2.9	61.8	79.4	82.4	85.3
RSICD	CLIP L/14	0.0	55.9	67.6	70.6	73.5

F Difficulty Stratification

Table 8: Success rates (%) by difficulty level. "Easy" experiments often have lower success than "hard" ones-embedding separability matters more than conceptual simplicity.

Difficulty	**UCM**	**AID**	**PatternNet**	**RSICD**
Easy	58.3	52.1	95.8	50.0
Medium	88.5	78.1	77.1	80.2
Hard	85.4	97.9	75.0	93.8
Very Hard	73.2	77.1	64.3	73.4

G Entanglement Threshold Sensitivity

Table 9: Success rates (%) by entanglement bin. Higher entanglement correlates with lower success.

Dataset	Low (0.3–0.5)	Med (0.5–0.7)	High (0.7–0.9)	Optimal τ
UCM	81.3	75.0	62.5	0.51
AID	75.0	70.8	54.2	0.82
PatternNet	79.2	72.9	58.3	0.65
RSICD	77.1	70.8	60.4	0.58
Mean	78.2	72.4	58.9	–

H Learned Composition Comparison

Table 10: Simple arithmetic vs TIRG-style learned composition. Arithmetic outperforms by 16–48 points even with 1,000–4,000+ training triplets.

Dataset	Model	Triplets	Arithmetic	TIRG
UCM	RemoteCLIP	1,260	**81.5**	33.3
AID	RemoteCLIP	2,610	**78.9**	31.6
AID	CLIP L/14	2,610	**78.9**	42.1
PatternNet	RemoteCLIP	4,218	**57.9**	42.1
PatternNet	CLIP L/14	4,218	**63.2**	26.3
Mean			**72.1**	35.1

I ROC Analysis

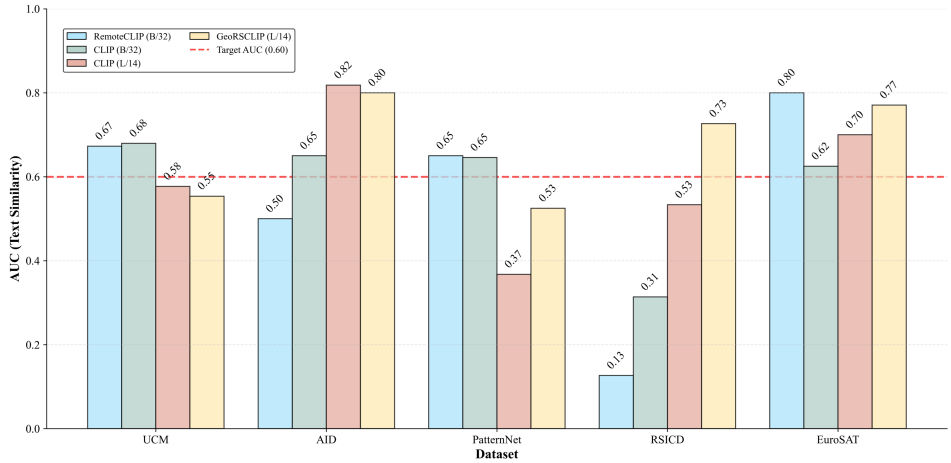

Figure 6: ROC curves for entanglement-based failure prediction across all model-dataset configurations.

The overall ROC comparison is presented in Figure 6. Per-dataset ROC curves for individual model-dataset configurations are available in the supplementary materials.

J SUMMARY STATISTICS

Table 11: Summary across all 4 models.

Model	Arithmetic	Prompt Var	AUC	Geometry
CLIP ViT-B/32	67.5	3.9	0.583	0.467
RemoteCLIP	65.9	6.1	0.550	0.646
CLIP ViT-L/14	72.8	13.1	0.599	0.443
GeoRSCLIP	73.8	4.6	0.675	0.590

K STATEMENT ON LLM USAGE

In accordance with ICLR's policy on the use of Large Language Models during paper preparation, we declare the extent and nature of LLM involvement in this work. We employed an LLM solely for text refinement purposes, including improving grammar, enhancing wording clarity, and rephrasing sections to meet academic writing standards. No scientific content, experimental results, or novel ideas were generated by the LLM.