

Semantic Instability in Vision-Language Classification of Abstract and Finished Sketches

Thomas Chen*

Independent Scholar

ORCID: <https://orcid.org/0009-0008-3869-9255>.

*Corresponding Author

Thomas Chen

Email:
thomasruntaochen@gmail.com

Article History

Received: 20.04.2026

Accepted: 04.06.2026

Published: 24.06.2026

Abstract: This study evaluates how OpenCLIP classifies abstract and finished sketches across synthetic and human-drawn corpora. Although vision-language models are now widely used for image retrieval and recognition, less is known about their response to simplified, hand-drawn stimuli that depart from photographic training regularities. Two matched datasets were tested: a 240-image synthetic corpus and a 120-image human corpus, each using 12 object categories and two drawing conditions. Each image was compared with a fixed 12-label text ensemble using zero-shot cosine similarity. Outcomes included Top-1 accuracy, Top-1-to-Top-2 margin, and confusion structure. Human drawings produced higher accuracy (95.0%) than synthetic inputs (87.1%), but synthetic inputs produced higher mean margins (0.081 vs 0.056). Finished drawings improved performance in both sources. Category-specific failures, especially synthetic plant and shoe, demonstrate that controlled abstraction can expose semantic instability and boundary conditions in vision-language image interpretation. The method offers a compact protocol for imaging-science evaluation of sketch robustness.

Keywords: *vision-language models; OpenCLIP; sketch recognition; semantic similarity; machine vision; abstraction; zero-shot classification; image interpretation.*

Cite this article:

Chen, T., (2026). Semantic Instability in Vision-Language Classification of Abstract and Finished Sketches. *ISAR Journal of Arts, Humanities and Social Sciences*, 4(6), 38-41.

Introduction

Vision-language models are increasingly used to associate images with text labels, retrieve visual content, and support machine-vision tasks without task-specific retraining. CLIP and related systems learn shared image-text embeddings in which an image can be compared directly with candidate textual descriptions, making zero-shot classification possible when no category-specific training set is supplied (Radford et al., 2021; Zhang et al., 2024). This capability is valuable for imaging science because it shifts evaluation from pixel-level restoration or enhancement alone toward questions of semantic interpretation: whether a model can assign a visually simplified input to the intended concept and how decisively it does so. Recent work on multimodal similarity also shows that semantic scores are model-, task-, and dataset-dependent, which means they require contextual interpretation rather than simple treatment as objective measures of meaning (Ihnaini et al., 2024; Li et al., 2024).

Sketches provide a useful test case for this problem because they preserve object identity through sparse contour, internal structure, and proportion rather than through photographic texture. They also represent a common class of non-photographic images found in education, design, diagrams, visual notes, and creative practice. Yet many benchmarked vision-language applications emphasize natural images or polished synthetic outputs, leaving less evidence about how these systems respond to controlled human abstraction. This gap is relevant to imaging-science readership because abstraction functions as a controlled degradation of semantic evidence: visual detail is reduced, but the target label remains known. A model that remains accurate under such reduction demonstrates robustness to sparse visual signals, while a model that drifts toward a competing label reveals a boundary condition in its learned representation.

The present study evaluates OpenCLIP classification of two matched sketch corpora. One corpus contains AI-mediated

synthetic sketch inputs expanded from original reference drawings; the second contains fully human-produced sketches using the same categories and conditions. Both corpora distinguish abstract and finished drawings, allowing source and visual detail to be examined as analytic factors. The study asks three questions. First, does recognition accuracy differ by source and drawing condition? Second, does decision margin, defined as the Top-1 cosine similarity minus the Top-2 cosine similarity, change differently from accuracy? Third, do errors cluster by category in ways that reveal semantic instability rather than random misclassification? The contribution is a compact protocol for testing sketch robustness in a pretrained vision-language model and a source-normalized analysis showing that accuracy and decisiveness can diverge.

Materials and methods

The comparative dataset contained 360 model-input images. The synthetic corpus contained 240 file-level inputs organized as 12 object categories, two drawing conditions, and 10 file-level repetitions per category-condition cell. The human corpus contained 120 student-produced drawings organized as the same 12 categories, the same two conditions, and five human variations per category-condition cell. The object categories were apple, block, books, bust, chair, hammer, keys, mug, plant, shoe, skull, and spoon. These labels remained fixed throughout evaluation and were the only candidate labels supplied to the model.

The two visual conditions were coded C1 and C2. C1 represented abstract sketches with reduced contour and limited internal detail. C2 represented finished sketches with clearer internal structure and more complete visual information. Example stimuli are shown in Figure 1. The synthetic corpus was expanded from one original reference drawing per category into abstract and finished variants. In some synthetic cells, fewer than 10 generated variants were available; duplicated files were used to complete the balanced file-level structure. This procedure balanced classification trials but did not create fully independent synthetic exemplars in every cell. Accordingly, the synthetic corpus is reported as file-level input data. The human corpus consisted of distinct human-produced drawings. All human images were anonymized, and no identifying information was included in filenames or image content.

Each image had one metadata row with file path, source, true label, condition, condition name, and repetition. Source was coded synthetic or human. The analytic output preserved source, true label, condition, predicted label, Top-1 similarity score, Top-2 label, Top-2 similarity score, margin, and correctness. Images were standardized as RGB inputs before model preprocessing. Category words or labels were not embedded in the images.

Evaluation used OpenCLIP ViT-B-32 with the laion2b_s34b_b79k checkpoint. This model family extends contrastive image-language learning in a reproducible open implementation, while preserving the zero-shot comparison logic established by CLIP (Cherti et al., 2023; Radford et al., 2021). No fine-tuning, calibration, or retraining was performed. For each of the 12 labels, five sketch-oriented text prompts were encoded: "a pencil sketch of a {label}", "a hand-drawn sketch of a {label}", "a simple line drawing of a {label}", "a black and white sketch of a {label}", and "a doodle of a {label}". The five normalized text embeddings were averaged and normalized again to produce one prototype vector per label. Each image embedding was normalized

and compared with all label prototypes using dot-product cosine similarity.

The primary outcome was Top-1 accuracy, coded as correct when the highest-scoring label matched the metadata label. The secondary outcome was margin, calculated as Top-1 similarity minus Top-2 similarity. Margin was interpreted as a decisiveness proxy rather than as a calibrated probability. Higher margins indicate stronger separation between the selected label and the nearest competing label, while lower margins indicate boundary ambiguity even when classification is correct. Results were summarized by source and condition, then by category within source and condition. Confusion matrices were used to determine whether errors formed interpretable pathways. Because the sources differed in sample size, comparisons emphasize rates, means, and category-level patterns rather than pooled raw counts. This avoids overstating the larger synthetic corpus and aligns with recent cautions that multimodal evaluation should combine quantitative measures with transparent interpretation (Hartwig et al., 2025; Otani et al., 2023).

Results

OpenCLIP classified the human corpus more accurately than the synthetic corpus, although the synthetic corpus produced more decisive similarity separations. Table 1 reports condition-level performance. In the synthetic corpus, Top-1 accuracy increased from 82.5% in C1 abstract images to 91.7% in C2 finished images. In the human corpus, accuracy increased from 93.3% to 96.7%. Overall accuracy was therefore higher for human drawings (95.0%) than for synthetic file-level inputs (87.1%). These results indicate that hand-drawn variation did not prevent recognition; indeed, the model selected the correct label more often for human images than for synthetic images.

Margin results changed the interpretation. Mean margin was higher for synthetic inputs in both conditions, increasing from 0.070 in C1 to 0.092 in C2. Human margins were lower, increasing from 0.050 to 0.063. Thus, the model was more often correct on human drawings, but its selected label usually sat closer to the nearest alternative. Conversely, the model was less accurate on synthetic inputs but made more separated label choices. Figures 2 and 3 visualize this divergence. The finding matters because accuracy alone would suggest that human sketches were simply easier for OpenCLIP, while margin shows that many human classifications remained closer to the decision boundary.

Category-level results showed that abstraction did not affect all categories equally (Table 2). Most synthetic categories were recognized perfectly in both conditions, including apple, block, books, bust, chair, keys, mug, skull, and spoon. The strongest synthetic failures were plant and shoe. Plant shifted from 0% accuracy in C1 to 100% in C2, indicating that added internal structure restored the correct label. Shoe remained at 0% in both synthetic conditions, indicating a persistent source-specific error. In the human corpus, errors were fewer and less extreme. Apple improved from 60% to 100%, plant from 80% to 100%, while hammer remained 80% in both conditions and chair declined from 100% to 80%.

Confusion matrices confirmed that errors were structured rather than random (Figure 4). In the synthetic abstract condition, all plant inputs were classified as spoon and all shoe inputs as bust, with one hammer input classified as plant. In the synthetic finished

condition, plant recovered completely, but shoe remained unstable, with most finished shoe inputs classified as skull. Human errors were more dispersed and often near ties. For example, a human abstract plant classified as mug retained plant as the second choice with a margin of 0.0008, and a finished hammer classified as spoon retained hammer as second choice with a margin of 0.0010. These cases show that semantic drift can appear either as a confident movement into a wrong category or as a near-boundary decision between competing labels.

Discussion

The results demonstrate that controlled sketch abstraction can expose semantic boundary conditions in a vision-language classifier. The clearest empirical pattern is the divergence between accuracy and margin. Human drawings produced the highest Top-1 accuracy, which suggests robust recognition of naturally variable hand-drawn stimuli. Synthetic inputs produced higher margins, which suggests that they often matched the model's learned image-text regularities more decisively. This does not mean that synthetic sketches were better images. It means that their visual structure created stronger embedding separation once the model selected a label. For imaging science, the distinction is important because an image-recognition system can be correct but unstable, or incorrect but confident. These two failure modes require different evaluation language.

The category-level findings further show that abstraction is not a uniform degradation. Some categories remained stable under visual reduction, while others crossed semantic boundaries. Synthetic plant illustrates recoverable instability: abstract inputs collapsed into spoon, but finished inputs reached perfect recognition. The likely explanation is that reduced plant contours lacked category-specific cues such as stems, leaves, and branching structure, while finished drawings restored them. Synthetic shoe illustrates non-recoverable instability within this dataset because added detail did not restore correct classification. Human plant and hammer errors were different; they often occurred with extremely small margins, indicating that the model had not decisively rejected the correct class. These patterns suggest two types of drift: confident misalignment, where the wrong label dominates, and boundary ambiguity, where the correct and incorrect labels nearly tie.

This study contributes a practical evaluation framework for non-photographic image interpretation. Rather than treating drawings as informal illustrations, it uses them as controlled stimuli for probing representation space. The approach complements broader work on multimodal evaluation, which has warned against relying on single metrics or under-described human judgments (Hartwig et al., 2025; Otani et al., 2023). Here, Top-1 accuracy, margin, and confusion structure provide complementary evidence. Accuracy measures categorical success; margin measures decisiveness; confusion matrices identify the direction of semantic substitution. The combination is especially useful for sketch, icon, diagram, and educational-image contexts, where sparse visual information may be meaningful to human viewers but unstable for machine interpretation.

The study also has implications for AI-assisted visual culture. Prior research on AI in creative practice emphasizes that artificial intelligence systems reshape how images are produced and interpreted, but much of that literature centers on generation rather than recognition (Oksanen et al., 2023). The present analysis shifts

attention to machine interpretation of human and AI-mediated visual forms. It shows that model error can be analytically useful because it reveals which visual cues are privileged by a learned embedding space. In settings where vision-language models label, retrieve, or sort cultural images, such boundary conditions deserve attention, particularly for stylized, simplified, or student-produced imagery.

Several limitations qualify the findings. The dataset is modest and limited to 12 object categories. The synthetic corpus includes duplicated file-level inputs in some cells, so variation within that source should not be interpreted as fully independent generation. The analysis uses one OpenCLIP architecture and one checkpoint; cross-model replication is needed before generalizing to vision-language systems broadly. The human corpus includes natural variation in drawing style, which improves ecological realism but reduces control over line quality and proportion. Finally, the label set was closed, so the model could choose only among 12 categories. Future work should test additional models, include formal human-viewer comparison, expand the category set, and use graded abstraction levels rather than a two-condition design. Even within these limits, the findings support the central claim: controlled abstraction provides a compact and reproducible method for detecting semantic instability in vision-language image interpretation.

Acknowledgments

The author gratefully acknowledges Dr. James Hutson, Department Head of Art History, Artificial Intelligence, and Visual Culture, for research mentorship and methodological guidance throughout the development of this project. Dr. Hutson provided support in research design, literature mapping, study conceptualization, dataset planning, evaluation methodology, and manuscript development. His mentorship helped shape the interdisciplinary integration of artificial intelligence, computer vision, visual culture, and artistic practice that informed this investigation.

Data availability: De-identified row-level results, analysis code, and metadata available upon reasonable request.

Funding: No external funding was received.

Declaration of interest: The author reports no competing interests.

References

1. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., & Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818-2829. <https://doi.org/10.1109/CVPR52729.2023.00276>
2. Hartwig, S., Engel, D., Sick, L., Kniesel, H., Payer, T., Poonam, P., Glockler, M., Baeuerle, A., & Ropinski, T. (2025). A survey on quality metrics for text-to-image generation. *IEEE Transactions on Visualization and Computer Graphics*, 31(10), 9464-9483. <https://doi.org/10.1109/TVCG.2025.3585077>
3. Ihnaini, B., Abuhaija, B., Atta Mills, E., & Mahmuddin, M. (2024). Semantic similarity on multimodal data: A comprehensive survey with applications. *Journal of King*

- Saud University: Computer and Information Sciences, 36(10), 102263. <https://doi.org/10.1016/j.jksuci.2024.102263>
4. Li, J., Kementchedjheva, Y., Fierro, C., & Sogaard, A. (2024). Do vision and language models share concepts? A vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12, 1232-1249. https://doi.org/10.1162/tacl_a_00698
 5. Oksanen, A., Cvetkovic, A., Akin, N., Latikka, R., Bergdahl, J., Chen, Y., & Savela, N. (2023). Artificial intelligence in fine arts: A systematic review of empirical research. *Computers in Human Behavior: Artificial Humans*, 1(2), 100004. <https://doi.org/10.1016/j.chbah.2023.100004>
 6. Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkila, J., & Satoh, S. (2023). Toward verifiable and reproducible human evaluation for text-to-image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14277-14286. <https://doi.org/10.1109/CVPR52729.2023.01372>
 7. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv*. <https://doi.org/10.48550/arXiv.2103.00020>
 8. Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8), 5625-5644. <https://doi.org/10.1109/TPAMI.2024.3369699>