TactileEval: A Step Towards Automated Fine-Grained Evaluation and Editing of Tactile Graphics

Adnan Khan[1,*], Abbas Akkasi[1] and Majid Komeili[1]

***Abstract*— Tactile graphics require careful expert validation before reaching blind and visually impaired (BVI) learners, yet existing datasets provide only coarse holistic quality ratings that offer no actionable repair signal. We present TACTILEEVAL, a three-stage pipeline that takes a first step toward automating this process. Drawing on expert free-text comments, we establish a five-category quality taxonomy, encompassing view match, required parts, identity & background, texture separation, and line quality, aligned with BANA standards. We subsequently gathered 14,095 structured annotations via Amazon Mechanical Turk, spanning 66 object classes organized into six distinct families. A reproducible ViT-L/14 feature probe trained on this data achieves 85.70% overall test accuracy across 30 different tasks, with consistent difficulty ordering suggesting the taxonomy captures meaningful perceptual structure. Building on these evaluations, we present a ViT-guided automated editing pipeline that routes classifier scores through family-specific prompt templates to produce targeted corrections via `gpt-image-1` image editing. Code, data, and models are available at `https://TactileEval.github.io/`.**

I. INTRODUCTION

Tactile graphics are embossed representations of visual content that allow BVI individuals to perceive images through touch. Producing high-quality tactile graphics is labor-intensive: experts must inspect line quality, texture encoding, posture, background clutter, and viewpoint consistency before embossed artifacts reach BVI learners [1], [2]. TactileNet [3] represents an important advance by curating natural/tactile pairs with expert quality ratings; however, those ratings are coarse: four holistic levels (Accept, Minor Edit, Major Edit, Reject). These do not specify *which* attributes are flawed nor *how* to correct them.

This paper presents TACTILEEVAL, a pipeline addressing this gap through three complementary contributions:

- **A fine-grained quality dataset** spanning 30 distinct task families, derived from the intersection of six object families and five BANA-aligned quality dimensions. The corpus includes 14,095 structured annotations across 66 TactileNet object classes, collected via a rigorous AMT protocol (Sections III–IV).
- **A reproducible ViT-based quality evaluator** that achieves 85.70% overall test accuracy across 30 different tasks, providing a structured alternative to expert-only annotation (Section V).
- **A ViT-guided editing pipeline** that translates automated quality scores into targeted image edits via family-specific prompt templates, taking a concrete step toward end-to-end tactile repair (Section VI).

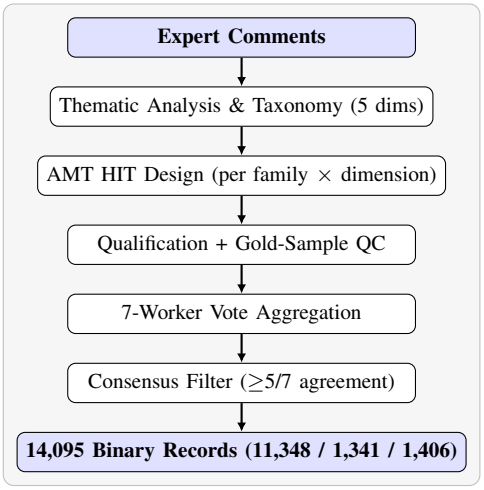


Fig. 1. Dataset construction pipeline. Expert free-text comments from TactileNet seed a five-category quality taxonomy, which drives AMT HIT design across all six object families. Qualified workers complete HITs with gold-sample quality control; votes are aggregated and filtered for consensus, yielding 14,095 binary records.

Fig. 1 illustrates the dataset construction stage. The pipeline deliberately decomposes expert tactile knowledge into structured, option-level judgments expressible in plain language; removing the reliance on domain specialists for each new annotation cycle and enabling scalable quality assessment across the full TactileNet catalog.

The remainder of this paper is organized as follows. Section II reviews related work; Section III develops the quality taxonomy; Section IV details dataset construction; Section V presents the ViT-based evaluator; Section VI introduces the editing pipeline; Sections VII–VIII discuss findings, limitations, and conclusions.

II. BACKGROUND AND RELATED WORK

A. TactileNet Dataset

TactileNet [3] is the first large-scale dataset for generating embossing-ready tactile graphics, pairing natural photographs across 66 object classes with tactile drawings at scale via Stable Diffusion models fine-tuned with LoRA and DreamBooth. Expert evaluation reported 92.86% adherence to tactile accessibility standards, with an SSIM of 0.538 between generated and expert-designed images. While TactileNet demonstrates tractable generation, its holistic quality

[1]School of Computer Science, Carleton University, Ottawa, Canada
[*]Corresponding author: Adnan Khan. E-mail: adnankhan5@cmail.carleton.ca

ratings neither identify which attributes fail nor prescribe corrections; the gap TACTILEEVAL addresses.

B. Tactile Graphics for BVI Learners

Tactile graphics are essential for conveying spatial information to BVI individuals [4], relying on raised lines, textures, and shapes to communicate visual structure through touch. Guidelines from BANA [5] and RNIB [6] codify best practices covering viewpoint simplification, texture separation, and minimum stroke weight; criteria that directly inform our five quality dimensions. Recent work has begun automating parts of this pipeline via rule-based vector simplification [7] and neural map-to-tactile translation [8]. On the evaluation side, however, automated quality assessment for tactile graphics remains comparatively underexplored: existing accessibility-focused systems largely rely on manual expert review or coarse holistic ratings rather than structured, attribute-level scoring, and no prior framework couples crowd-derived quality taxonomies with a trainable, reproducible evaluator. This gap, distinct from the generation-side advances above, is the central focus of TACTILEEVAL.

C. Crowdsourced Quality Assessment

Crowdsourced annotation via Amazon Mechanical Turk has been widely adopted for perceptual image quality benchmarks [9], [10]. Crowd labels can match expert-level quality when majority-vote aggregation and worker filtering are applied [11]; principles we adopt directly. Tactile graphics introduce additional requirements: workers must compare a natural photograph with a line-art drawing, demanding targeted training and qualification gating beyond standard image quality tasks.

D. Vision-Language Models for Visual Assessment

Contrastive vision-language models such as CLIP [12] build on contrastive self-supervised learning [13] to produce features that transfer to downstream classification via lightweight probing. MLLMs including GPT-4o [14] and LLaVA [15] achieve strong performance on visual question answering through instruction tuning. No prior work applies either paradigm to tactile or accessibility-specific quality assessment.

III. QUALITY TAXONOMY

A. Mining TactileNet Expert Comments

TactileNet's annotation form included an optional free-text field describing perceptual problems in each tactile drawing. We performed a qualitative thematic analysis of approximately 80 non-empty expert comments from "non-accept" image pairs to surface recurring tactile failures. This analysis formed the basis for our structured taxonomy.

B. Five BANA-Aligned Quality Dimensions

We established five quality dimensions grounded in BANA standards: **View Match** (V), **Required Parts** (P), **Identity & Background** (B), **Texture Separation** (T), and **Line Quality** (L). We mapped these dimensions across six object families (Animals, Food/Nature, Vehicles, etc.), resulting in 30 distinct task families (F*n*Q*X*). To ensure high-quality crowdsourced labels, we adapted the instructions for each family to use domain-specific terminology. For example, the **Required Parts (P)** prompt for Animal families (F1) focuses on anatomical features like limbs and tails, whereas for Simple Objects (F2) it targets functional components like stems or handles. Each task is a multi-select probe requiring workers to identify specific failure modes:

- **View Match (QV):** Compares tactile orientation (e.g., Front, Side, Top) against the reference photo to prevent perspective-driven confusion.
- **Required Parts (QP):** Identifies missing, hallucinated, or incorrectly placed components relative to the object's canonical structure.
- **Identity & Background (QB):** Flags background clutter, extraneous artifacts, or failures in preserving the object's semantic class (e.g., mismatched species or vehicle type).
- **Texture Separation (QT):** Evaluates tactile differentiation between adjacent regions, flagging inconsistent patterns, boundary crossings, or excessive density.
- **Line Quality (QL):** Assesses the physical integrity of the drawing's outlines, specifically identifying broken, fuzzy, or colliding bold strokes.

TABLE I
OPERATIONALIZATION OF QUALITY DIMENSIONS ACROSS FAMILIES

Dimension	Animal (F1) Context	Food/Objects (F2) Context
Parts (P)	Missing limbs, tails, ears	Missing stems, leaves, handles
Texture (T)	Fur vs. wing differentiation	Skin vs. seed differentiation
Identity (B)	Species/breed accuracy	Type (e.g., fruit vs. tool) accuracy
Lines (L)	Solid, traceable contours	Solid, traceable contours

Table I demonstrates how these generic dimensions were operationalized using family-specific prompts. Table II lists all checkbox options per quality dimension. Each dimension contains between 3 and 7 options; F*n*QT is the most granular, reflecting the subjective nature of texture judgments.

IV. DATASET CONSTRUCTION

A. AMT Data Collection Protocol

Each Human Intelligence Task (HIT) presented workers with a side-by-side view of a natural image and its corresponding tactile drawing. Workers selected all applicable failure options from a dimension-specific checkbox list, or indicated no issue if the tactile was satisfactory. Annotation quality was enforced through a two-stage vetting process.

a) Training examples.: Each task's instruction page presented three annotated reference pairs representing the full spectrum: a clean tactile, a single clear defect, and a pair with multiple co-occurring issues.

TABLE II

CHECKBOX OPTIONS PER QUALITY DIMENSION AND THEIR FOCUS

Dim.	What workers assess
QB	Species/object identity, pose or configuration match, background cleanliness
QL	Broken or discontinuous outlines, overly bold strokes, blurry/low-contrast lines, or no issues
QP	Missing anatomical parts, hallucinated structures, incorrect placement, or all correct
QT	Absent texture, intra-element inconsistency, near/far confusion, region bleed, density, similarity to adjacent regions, or no issues
QV	View-angle match; if mismatch: frontal, side, top, or perspective; F2 adds an option for geometrically ambiguous viewpoints

b) Gold-sample quality control.: Each main HIT embedded at least two gold-sample questions with manually verified labels. Assignments failing the gold threshold were rejected and re-posted to maintain annotation quality.

c) Qualification test.: A custom qualification test per task family required workers to meet a 2/3 accuracy threshold; only one attempt was permitted per 24-hour window. Task visibility was set to *Private*: only qualified workers could accept assignments.

d) Label aggregation.: For each image pair and checkbox option, we compute the vote fraction across 7 workers. An option is labeled TRUE when the vote fraction meets or exceeds a per-task confidence threshold: ≥ 0.6 for F*n*QV, F*n*QP, F*n*QB, and F*n*QL—i.e. at least 5 out of 7 workers in agreement, a conservative super-majority that goes beyond the strict 4/7 (≈ 0.571) majority cutoff—and > 0.4 for F*n*QT, where texture judgments exhibited higher inter-worker variability and the stricter threshold suppressed valid positive annotations.

B. Full-Scale Dataset

We organise the 66 TactileNet object classes into six broad families: Animals & Creatures (F1), Food, Nature & Simple Objects (F2), Furniture & Structures (F3), Tools, Instruments & Appliances (F4), Vehicles & Flight Systems (F5), and Wearables & Accessories (F6). The AMT collection was deployed across all six families and all five quality dimensions, yielding 30 distinct task families.

Raw AMT results were reprocessed across all families with normalized option schemas and a two-stage consensus filter: approved ballots are kept; additionally, votes showing agreement among at least five workers on an identical label vector are promoted, while under-supported or tied options are dropped. The resulting dataset contains only consensus-backed binary decisions along with exact vote counts per pair and option.

The final dataset comprises **14,095** option-level binary records split into **11,348** training / **1,341** validation / **1,406** test examples. Each record contains the image pair identifiers, task family, option identifier and description, majority label, vote fraction, vote counts, and provenance metadata. Because each image pair is evaluated across multiple options (one binary record per option), the dataset captures co-occurring defects within a single structured representation.

The F1QV prompt for Animals & Creatures illustrates the plain-language design:

> **F1QV prompt:** "Does the animal show the same flat view as the photo (same head/body angle)?"
>
> **Options:** `angle_match` · `view_frontal` · `view_side` · `top_view` · `view_perspective`

V. VIT-BASED QUALITY EVALUATION

A. Feature Probe Architecture

To provide a reproducible, cost-transparent evaluator, we train a ViT-based feature probe that deliberately separates representation from decision-making.

a) Feature extraction.: For each record, we extract image embeddings from both the natural and tactile images using a frozen CLIP ViT-L/14 model [12] initialized from the `laion2b_s34b_b88k` pretrained checkpoint [16]. A text embedding is extracted from the option prompt *"Task {family} option {option_id}: {description}"*. All embeddings are ℓ_2-normalized; each image and text embedding is 768-dimensional, yielding a concatenated feature vector of $4 \times 768 = 3{,}072$ dimensions. The final vector concatenates: the natural image embedding, the tactile image embedding, their element-wise difference (capturing cross-modal discrepancy), and the text option embedding.

b) Classifier and training.: Each quality option is treated as an independent binary classification problem (issue present vs. absent), with the sigmoid output serving directly as an issue probability. A softmax multi-class formulation over all options within a task family was also explored but yielded sub-optimal results, likely because options are not mutually exclusive and co-occurring defects are common. A two-layer MLP (Linear(3072, 512)–ReLU–Linear(512, 1)) is trained per option using binary cross-entropy loss and AdamW (lr 1×10^{-3}) with batch size 128 over 20 epochs, selecting the best checkpoint by peak validation accuracy. The CLIP ViT-L/14 backbone is fully frozen throughout; only the MLP weights are updated. This design choice is deliberate rather than a simplification: a frozen backbone keeps the evaluator fully reproducible on a single consumer GPU, avoids overfitting the pretrained representation to our comparatively small option-level corpus, and isolates the question of whether tactile quality is linearly recoverable from existing vision-language embeddings, a property with direct implications for transferability to future object families and quality dimensions without backbone retraining. We view end-to-end fine-tuning or cross-attention fusion between natural and tactile embeddings as a natural extension once a larger annotation corpus removes the overfitting risk; the present architecture establishes a reproducible, cost-transparent floor against which such extensions can be mea-

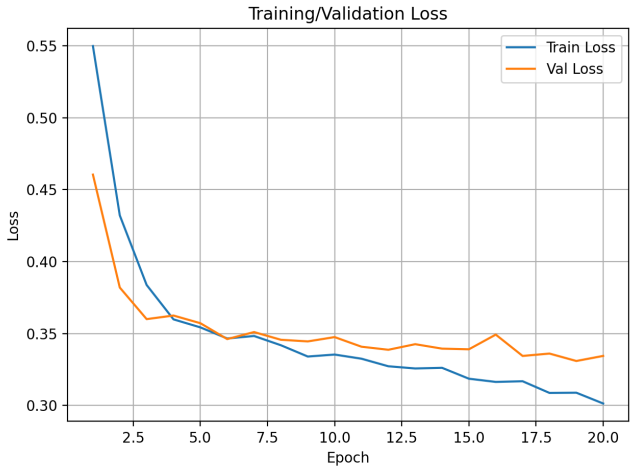


Fig. 2. Training and validation loss over 20 epochs. Both curves converge smoothly with no divergence, indicating stable generalisation throughout training.

sured. All experiments run on a single NVIDIA RTX 4090 (24 GB VRAM).

c) Training dynamics.: Fig. 2 shows the training and validation loss curves. Both losses drop sharply in the first three epochs and converge smoothly thereafter, with no sign of overfitting. The slight early-epoch lead of validation loss over training loss is an artefact of within-epoch averaging: training loss is computed cumulatively as weights update across mini-batches, so early batches are scored against a less-converged model, while validation loss is computed once per epoch using the fully updated weights. This pattern disappears once both curves stabilise, and is consistent with the deterministic, pair-level train/validation/test partitioning described in Section IV-B, which guarantees no natural-tactile pair appears in more than one split. We note that all splits are pair-level rather than class-level; a class-level holdout, in which entire object categories are withheld from training, would provide a stronger test of cross-category generalisation and is a natural direction for future validation.

B. Evaluation Results

Table III reports per-family test accuracy across all six object families. The probe achieves **85.70%** overall test accuracy on 1,406 test records. Family-level accuracy ranges from 79.51% (F2, Food, Nature & Simple Objects) to 89.31% (F5, Vehicles & Flight Systems).

Fig. 3 visualizes accuracy across all 30 task families. Tasks such as F5QB, F6QB, and F4QB (identity/background checks) reach perfect accuracy, consistent with background cleanliness being a visually unambiguous signal. Challenging tasks concentrate in F1QP and F4QP (required parts), where minimalistic tactile drawings can legitimately omit detail, making the part-presence judgment inherently subjective.

At the option level (Fig. 4), the hardest cases include `missing_parts` (0.50) and `missing_texture` (0.68), requiring localized, fine-grained reasoning about fill density or anatomical completeness that global ViT

TABLE III

ViT FEATURE PROBE TEST ACCURACY ACROSS ALL SIX FAMILIES

Family	Object Category	Acc. (%)
F1	Animals & Creatures	84.72
F2	Food, Nature & Simple Objects	79.51
F3	Furniture & Structures	87.25
F4	Tools, Instruments & Appliances	80.00
F5	Vehicles & Flight Systems	89.31
F6	Wearables & Accessories	87.86
Overall		**85.70**

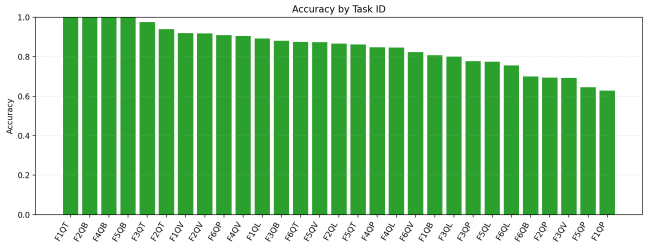


Fig. 3. ViT feature probe test-set accuracy across all 30 tasks (sorted descending). Identity/background checks (F5QB, F6QB, F4QB) reach 1.0, while required-parts checks (F1QP, F4QP) prove the most challenging.

embeddings struggle to resolve. Geometric and identity checks such as `angle_match`, `bleeds_boundaries`, and `configuration_match` reach 1.0, indicating that the difficulty ordering reflects genuine task structure rather than annotation artefacts.

VI. ViT-Guided Editing Pipeline

While the ViT evaluator identifies *what* may be wrong with a tactile graphic, correcting identified defects requires translating classifier scores into targeted image edits. We present a ViT-guided editing pipeline as a concrete step in this direction, with the explicit acknowledgment that a comprehensive tactile correction system remains an open challenge. The case studies in Section VI-C focus on single-issue correction to isolate defect-specific quality; the underlying pipeline supports multi-issue prompts directly via a configurable parameter.

A. Pipeline Design

For a given tactile image and target task family, the pipeline proceeds as follows (Fig. 5).

a) Issue scoring.: The natural image, tactile image, their element-wise embedding difference, and each option's text prompt are encoded by the frozen ViT-L/14 backbone and the trained MLP head. The resulting issue probability inverts the raw sigmoid output for negative-polarity pass signals (e.g., `no_line_issues`), so that all scores express the probability of a defect. Pass-type options are additionally excluded from the candidate pool by an explicit non-actionable filter, preventing them from ever competing as repair targets regardless of confidence. Among the remaining actionable options, every option whose issue probability

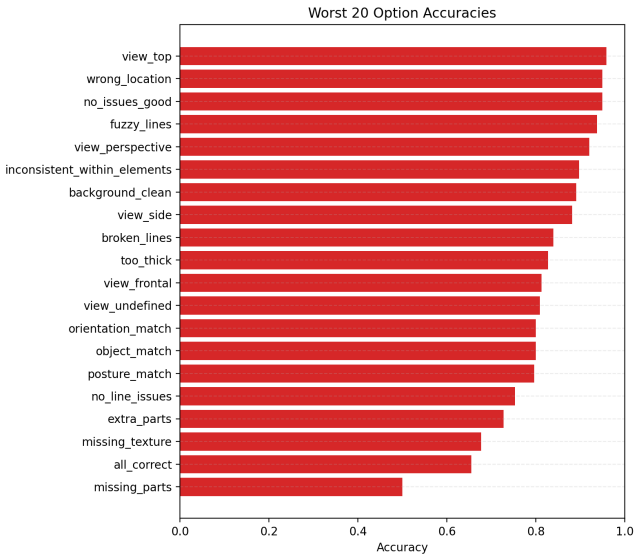


Fig. 4. Bottom-20 option accuracies (per-option accuracy is the fraction of correctly predicted samples for that option independently; options are not macro-averaged).

meets or exceeds a fixed threshold (0.6) is selected, naturally accommodating co-occurring defects rather than committing to a single diagnosis; if no option clears the threshold, the highest-scoring actionable option is selected as a fallback. The number of issues forwarded to the prompt is governed by a configurable issues-per-edit parameter, set to one for the case studies in Section VI-C to isolate single-defect correction quality, though the underlying selection mechanism supports multi-issue prompts directly.

b) Prompt construction.: The selected issue code is mapped to a natural-language repair instruction via `ISSUE_TEMPLATES`. This is assembled into a structured prompt with a family-level header and footer from `FAMILY_PROMPT_FRAMES`, which enforce tactile formatting constraints (clean silhouettes, stroke continuity, background cleanliness) as guardrails independent of the specific defect.

c) Image editing.: The original tactile is padded to a square canvas and submitted as the base image to `gpt-image-1` in edit mode, so edits are applied to the existing drawing rather than generating from scratch. The pipeline saves the prompt, edited image, and structured metadata for human review.

B. Zero-Shot Baseline

To motivate the structured ViT-driven approach, we first evaluated two unguided generation modes. **Text-only** conditioning produced unreliable species identity: the Penguin output (Fig. 6a) resembles a marine mammal despite explicit specification. **Natural-conditioned** generation improved species fidelity but consistently reintroduced background habitat context (Fig. 6b). Both modes required expert triage before embosser use, confirming that unstructured generation alone is insufficient.

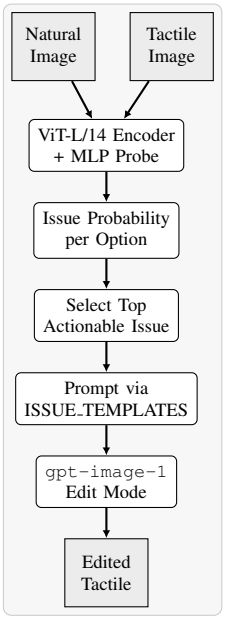
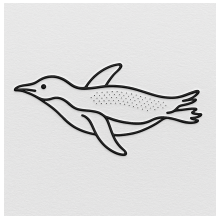
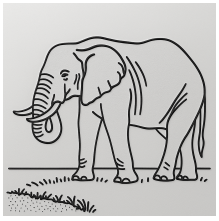


Fig. 5. ViT-guided editing pipeline. The natural and tactile images are encoded by the frozen ViT-L/14 backbone; the MLP probe assigns issue probabilities per option. The top actionable issue maps to a family-specific repair instruction that is submitted to `gpt-image-1` in edit mode, constraining corrections to the existing drawing.



(a) Text-only (Penguin)



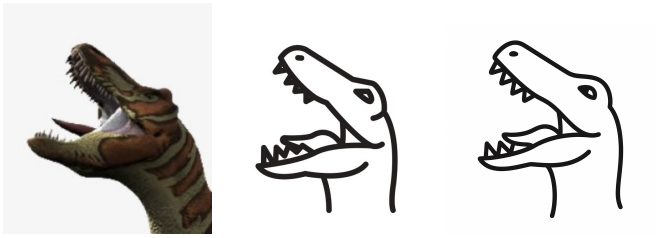
(b) Natural-conditioned (Elephant)

Fig. 6. Zero-shot outputs fail without structured issue guidance: (a) species fidelity breaks; (b) background clutter persists despite explicit removal instructions.

C. Editing Case Studies

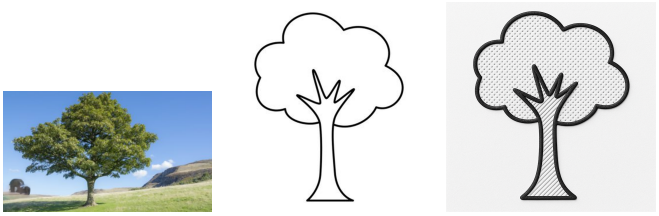
Fig. 7 illustrates the pipeline on a Dinosaur pair from Family 1 (Animals & Creatures), Line Quality (F1QL). AMT workers were unanimous (7/7) in flagging `too_thick`; the ViT probe likewise assigned a 0.985 issue probability, so the prompt emphasized thinning oversized strokes and separating the teeth into clean, distinct contours. The ViT-guided edit (Fig. 7c) clearly reduces the black fills in the mouth region and restores the outline legibility. Interestingly, the post-edit ViT score decreases only marginally ($\Delta p = -0.004$, see Section VI-C.1), highlighting that small perturbations near the decision boundary do not necessarily match human perception; the qualitative improvement is obvious even if the classifier's confidence barely changes. This underscores why we pair quantitative proxies with visual inspection in the editing study.

The Tree case (F2QT) showcases the pipeline on a texture-separation defect. AMT workers (7/7) labeled the tree as `missing_texture`, and the ViT probe assigned a 0.699



(a) Natural Photo (b) Original Tactile (c) ViT Edit

Fig. 7. F1QL editing case study: Dinosaur. (a) Natural photo reference. (b) Original tactile: excessively thick strokes in teeth. (c) ViT-guided edit (ViT issue prob. 0.985, `too_thick`).



(a) Natural Photo (b) Original Tactile (c) ViT Edit

Fig. 8. F2QT editing case study: Tree (missing texture). (a) Natural photo reference. (b) Original tactile: uniform texture fails to separate canopy regions. (c) ViT-guided edit (issue prob. 0.699→0.121) introduces differentiated fills and clear negative space.

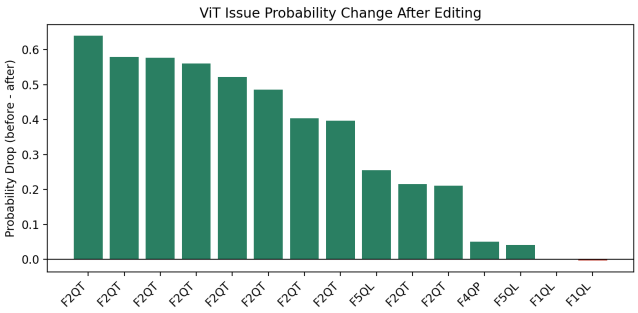


Fig. 9. Drop in ViT issue probability (before minus after) for the 15 high-confidence edits. Positive values indicate improvement; only one sample (F1QL dinosaur) slightly worsened.

TABLE IV

BEFORE/AFTER VIT PROBABILITIES ON SELECTED EVALUATION PAIRS.

Task / Option	p_{before}	p_{after}	Δ
F2QT Egg – `missing_texture`	0.903	0.693	+0.211
F2QT Planet – `missing_texture`	0.739	0.099	+0.640
F2QT Tree – `missing_texture`	0.699	0.121	+0.577
F5QL Scooty – `too_thick`	0.934	0.679	+0.255
F4QP Laptop – `missing_parts`	0.858	0.807	+0.050
F1QL Dinosaur – `too_thick`	0.985	0.989	−0.004

issue probability. After editing, the model introduces high-contrast fills for the canopy and trunk, carving out negative space that makes each region tactilely distinct. The ViT score drops to 0.121 ($\Delta p = 0.577$), aligning with the visual improvement and demonstrating that texture-focused prompts can yield precise, localized edits.

1) Quantitative proxy.: To move beyond single-case anecdotes, we selected 15 high-confidence crowd-supported defects in the test split (all with ≥ 5 worker votes and ViT issue probability ≥ 0.6, i.e. above the consensus threshold of Section IV-A) and ran the editing pipeline on each. We then re-scored the edited tactiles with the *same* ViT probe (no further fine-tuning) to measure the change in issue probability. Fig. 9 shows the per-sample deltas: 14/15 edits reduced the ViT signal (mean drop 0.329, median 0.397), indicating that the generated fixes move the render closer to the "no defect" manifold. Table IV lists representative pairs; note that the only regression occurred on the F1QL dinosaur example, matching the visual failure mode discussed earlier. We emphasize that re-scoring edits with the same probe used for diagnosis is a proxy rather than an independent measure of quality; incorporating sketch-specialized perceptual metrics (e.g., FID computed on sketch-domain features, or edge-direction histograms relative to expert-corrected references) alongside the BVI validation in Section VII-D would close this gap and is the most direct way to decouple editing success from the diagnosing classifier itself.

D. Cross-Family Editing Patterns

Across further experiments spanning all five quality dimensions, consistent patterns emerge. In F1QP (Required Parts), high-confidence ViT detections of `missing_parts` yield stable anatomical completions; both limb sets and head restored, while lower-confidence detections sometimes under-correct. In F1QT (Texture Separation), the pipeline's weakest dimension in evaluation (F1QT `missing_texture`: 0.68), the editing model tends to over-texture the full canvas rather than targeting specific body regions, consistent with the ViT's known difficulty in localizing fill-density defects. In F1QB (Identity & Background) and F1QV (View Match), where ViT accuracy is higher, the family-level guardrails in the prompt footer provide implicit normalization beyond the specific repair instruction; background cleanliness and posture alignment improve even when not the primary target. These patterns suggest that ViT-guided editing is most reliable when classifier confidence is high and the defect is spatially localized, and that the per-family prompt templates provide a useful floor of correction quality even under imperfect diagnosis.

VII. DISCUSSION

A. Structured Taxonomy as a Scalability Lever

A key design choice in TACTILEEVAL is decomposing expert tactile knowledge into plain-language, option-level judgments. This allows non-specialist crowd workers to contribute reliably at scale, removing the bottleneck of requiring a tactile accessibility expert for every new annotation cycle. The 85.70% overall ViT accuracy across 30 task families confirms that this structured signal is learnable from crowd labels alone, and the consistent difficulty ordering (geometric checks easiest, fine-grained texture and part checks hardest) reflects genuine perceptual structure rather than annotation noise.

B. ViT Probe as Evaluator and Editing Signal

The probe's role as an *editing signal* is distinct from its role as a classifier. Even when the ViT-selected issue code diverges from the dominant crowd judgment, the resulting edit still improves the tactile, because the family-level formatting guardrails in the prompt template apply a consistent floor of correction quality. This suggests that issue attribution granularity matters less than prompt quality for the editing step.

The probe does exhibit systematic probability inflation on high-frequency training options; notably `missing_texture` in F1QT and `missing_parts` across families—leading the ViT channel to over-generalize toward default repairs rather than instance-specific ones. Per-option confidence threshold calibration using held-out validation data is the natural mitigation.

C. Limitations

First, the `gpt-image-1` API dependency limits full reproducibility of the editing step; the ViT probe, dataset, and AMT protocol are fully open. Benchmarking the editing stage against open-source, line-art-specialized alternatives (e.g., ControlNet-conditioned diffusion or InstructPix2Pix) would yield a fully self-contained pipeline and is a concrete direction we intend to pursue. Second, the editing pipeline has not yet been evaluated by BVI participants or domain tactile transcribers; all reported gains rely on the ViT probe as a proxy for perceptual quality rather than direct end-user feedback. The qualitative results in Section VI-C are encouraging, but are insufficient to claim embosser-readiness.

D. Future Work

Per-option confidence threshold calibration using held-out validation splits per family is a natural next step to address probability inflation on high-frequency options. Extending the ViT probe with a lightweight decoder to produce structured rationales would enable self-contained diagnosis. A study with BVI participants remains an important direction to validate whether pipeline-corrected graphics yield better tactile comprehension outcomes in practice, complementing the proxy-based validation presented in this work.

VIII. Conclusion

We presented TactileEval, a three-stage pipeline for fine-grained evaluation and automated editing of tactile graphics. A five-category BANA-aligned taxonomy operationalizes expert knowledge as plain-language crowd tasks, enabling a 14,095-record dataset that spans all 66 TactileNet object classes (consolidated into six families) and is collected at scale without requiring specialist annotators. A ViT-L/14 feature probe trained on this data achieves 85.70% overall test accuracy across 30 task families, with difficulty ordering consistent across families. A ViT-guided editing pipeline translates probe outputs into targeted `gpt-image-1` edits, producing improved tactile renders in high-confidence, spatially localized defect cases.

We release all data, code, and model configurations and hope this work provides a concrete foundation for future BVI-validated, fully automated tactile correction pipelines.

Acknowledgment

This work was supported in part by MITACS and the Digital Research Alliance of Canada. The authors thank the student volunteers at the Intelligent Machines Lab (iML), Carleton University, for their contributions, Joshua Olojede and Hoda Vafaeesefat for their help with the AMT annotation environment.

References

[1] M. Mukhiddinov and S.-Y. Kim, "A systematic literature review on the automatic creation of tactile graphics for the blind and visually impaired," *Processes*, vol. 9, no. 10, p. 1726, 2021.

[2] P. Červenka, M. Hanousková, L. Másilko, O. Nečas, *et al.*, "Tactile graphics production and its principles," *Brno: Masaryk University Teiresiás–Support Centre for Students with Special Needs*, 2013.

[3] A. Khan, A. Choubineh, M. A. Shaaban, A. Akkasi, and M. Komeili, "Tactilenet: Bridging the accessibility gap with ai-generated tactile graphics for individuals with vision impairment," in *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 569–576, IEEE, 2025.

[4] P. Edman, *Tactile graphics*. American Foundation for the Blind, 1992.

[5] Braille Authority of North America, "Guidelines and standards for tactile graphics," 2010. Available at `http://www.brailleauthority.org`.

[6] Royal National Institute of Blind People, "Making diagrams accessible to blind and partially sighted people," 2010. RNIB Technical Guidance.

[7] T. Magne and O. Sorkine-Hornung, "Single-line drawing vectorization," in *Computer Graphics Forum*, vol. 44, p. e70228, Wiley Online Library, 2025.

[8] A. Choubineh, A. Akkasi, A. Khan, and M. Komeili, "Text-guided image-to-image translation for tactile map generation," in *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, IEEE, 2025.

[9] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015.

[10] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3, IEEE, 2019.

[11] R. Snow, B. O'connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks," in *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263, 2008.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[13] A. Khan, S. AlBarri, and M. A. Manzoor, "Contrastive self-supervised learning: a survey on different architectures," in *2022 2nd international conference on artificial intelligence (icai)*, pp. 1–6, IEEE, 2022.

[14] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[15] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023.

[16] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: an open large-scale dataset for training next generation image-text models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, (Red Hook, NY, USA), Curran Associates Inc., 2022.