PERCEPTUAL REPORT TRANSLATION: WHAT MUST AN ILLUSTRATION PRESERVE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image models can illustrate a perceptual account, but preserving its objects does not preserve how they were experienced. We introduce Perceptual Report Translation (PRT), a source-linked workflow and atlas, and examine the path from reports to representations to illustrations. Across 3,145 Ganzflicker records, narrative encoders predict complex content well (AUROC 0.887–0.894), while separate imagery/context responses improve access to vividness. A held-out retrieval audit shows that nearest narratives transfer content more reliably than vividness, persistence or affect. In 1,344 synthetic comparisons per encoder, counterbalanced wording and supervised probes separate similarity failure from recoverable distinctions. A 24-report image pilot finds high visible-detail coverage without a demonstrated structured-prompt advantage. Removing colour reverses all 45 eligible judge answers, yet none of the 69 original questions directly tests memory, uncertainty or analogy status. The atlas makes these gaps inspectable by retaining accounts, questionnaire responses, evidence and illustrations together. The results identify concrete requirements for perceptual communication: retain separate reporting channels, evaluate the deployed readout, and distinguish visible-content accuracy from preservation of the speaker's meaning.

1 INTRODUCTION

For someone living with a mental illness or neurological disorder, describing an unusual perceptual experience can be difficult even when it is vivid. An illustration might help a family member or clinician understand a visual disturbance. The aim is communication: give another person something concrete to discuss while preserving the account's qualifications and context. A diagnosis alone cannot specify what an individual sees or how the experience feels.

Figure 1 makes the problem concrete. A contributor describes a *memory* of a beach. A generated beach can match the nouns perfectly while leaving a reader unsure whether it depicts a memory, imagination, or something the person saw. We call this distinction the account's *claim status*: how the speaker presents the content, including uncertainty and analogy. Communicating the content and communicating its status are separate requirements.

A useful bridge must preserve more than resemblance. The narrative determines which information is available; the representation and its readout determine what a system can retrieve; the scoring criterion determines what counts as a successful illustration. We introduce *Perceptual Report Translation* (PRT), an implemented report-to-illustration workflow and descriptive atlas that keep these stages linked to their source.

Using 3,145 public Ganzflicker records (Reeder, 2022), we study three questions. First, which questionnaire qualities can narrative models predict, and which transfer to a similar real report? Second, when does wording obscure a qualification that a trained readout can recover? Third, does a high illustration score reflect the intended source meaning and sensitivity to visible evidence? The results connect a real-data representation study, controlled text diagnostics, and a 24-report image comparison. Clinical communication motivates the application; the induced-experience data provide an accessible testbed rather than clinical validation.

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.

a) The source account

"It was mostly visual memories of a sunny beach in my childhood."

Account 1318 · Ganzflicker

b) What the record retains

Visible content: beach
Status: memory
Source: exact quoted span

The score asks about the beach. It does not test memory status.

c) The PRT illustration



Figure 1: **Recognizable content does not specify how it was experienced.** (a) An actual Ganzflicker account; (b) selected information retained in its frozen evidence record; (c) its earliest structured-PRT illustration. The original evaluation asks whether a beach is visible. A yes answer cannot establish that a reader understands this as a memory. The image illustrates the report; it is not a participant-endorsed reconstruction.

The original studies are supplemented by exploratory threshold/calibration sensitivity, real-report retrieval, counterbalanced wording tests and a colour intervention. These additions were designed after inspecting earlier results; their protocols, complete outputs and timing are retained (Appendix M).

2 PRT: PRESERVE THE ACCOUNT THROUGH THE TRANSFORMATION

An implemented translation system. PRT has three components: a source-linked evidence record, a deterministic prompt compiler, and an atlas that exposes the account, questionnaire responses and generated outputs together. Its generator is an existing image model. Given account r, each annotation $e_j = (s_j, v_j, t_j)$ contains an exact source span s_j, an interpretation for illustration v_j, and a claim-status tag t_j. The compiler checks that s_j occurs in r, rejects missing evidence, and includes the full account in its request. Exact-span validation establishes provenance; interpretation remains an annotation decision. All annotations here were AI-authored before generation.

What the system preserves. Consider report 1899: "Lots of spirals and billowing movements, like when you pour milk into coffee." The record separates visible swirls from the coffee analogy; its prompt asks for the appearance without silently treating a drink as an observed object. For report 5403, "dark areas" becomes a visible-detail instruction while changing speed and counter-clockwise motion remain in the accompanying account. For the beach memory in Figure 1, the picture can illustrate remembered content while the record preserves how it was described. The shared policy retains qualifications and excludes diagnosis-based stereotypes. The matched study always requests an image; these instructions do not guarantee the output's fidelity.

The atlas as a research artifact. The offline atlas contains 3,145 searchable accounts, questionnaire filters and side-by-side report comparisons. Its illustrated subset links all 84 outputs to 24 source accounts, with prompts, statuses and image provenance available for inspection. This makes an interpretation contestable: a reader can return from an image or retrieved neighbour to the actual wording. It is a descriptive collection and working interface, not a diagnostic taxonomy. Figure 7 shows the implemented view; Appendix F documents the schema.

Separating evidence from its presentation. Let r be the account, e its fixed evidence record, π the shared preservation policy, and G the image backend. We compare

$$\begin{aligned} I_D &\sim G(\text{task}, r), \\ I_E &\sim G(\text{task}, \pi, \text{prose}(e), r), \\ I_S &\sim G(\text{task}, \pi, \text{fields}(e), r). \end{aligned} \tag{1}$$

Every condition includes the verbatim account. Evidence prose (E) and structured PRT (S) include the same statuses, spans, and interpretations; only presentation differs. Comparing S with direct prompting (D) instead measures the combined effect of annotation, instructions, and structure. These matched inputs separate the effect of presentation from the combined effect of supplying annotations and policy.

3 Related work

Ganzflicker studies connect imagery capacity with induced experience (Reeder, 2022; Chkhaidze et al., 2026); MOSAIC analyzes stroboscopic accounts with topics and language models (Beauté et al., 2026). MUSE maps foreground embodied sensory experience (Melvin et al., 2021), and low-vision simulations distinguish plausible transformations from intended perception (Natalie et al., 2025). These establish the relevance of perceptual accounts without supplying participant-endorsed image targets here.

Sentence encoders (Reimers & Gurevych, 2019; Song et al., 2020) and CLIP (Radford et al., 2021) supply reusable representations. Compositionality tests expose failures beyond object agreement (Yuksekgonul et al., 2022; Quantmeyer et al., 2024), while benchmark shortcuts (Hsieh et al., 2023) and the distinction between encoding and readout (Koishigarina et al., 2026; Lu et al., 2026) motivate our paired diagnostics and probes.

Image hallucination (Lim et al., 2025), semantic uncertainty (Franchi et al., 2025), metric sensitivity (Ross et al., 2024), and figurative grounding (Wang et al., 2026; Chen et al., 2026) are established evaluation problems. Our application links them to source qualifications in first-person perceptual reports. We test what our representations and criterion preserve, with source-linked artifacts that expose the remaining communication questions.

4 Study 1: which reported qualities are recoverable?

A narrative naming a geometric pattern may say little about how vivid, persistent, or unpleasant it was. We test recovery of these qualities using separate questionnaire responses, rather than model-invented labels.

Data and evaluation. We retain consented Ganzflicker records with reported ages 18–100 and descriptions of 5–400 words: 3,145 records, comprising 3,134 normalized-text groups. Targets distinguish complex objects/environments, at least moderate vividness, persistence of several seconds or longer, and unpleasant affect. Ambiguous or missing targets are excluded per task. These are heterogeneous self-reports without a validated person identifier; records are not necessarily distinct people. Appendix A gives exact eligibility and label mappings.

Text-hash groups define approximately 60/20/20 training/development/test splits (1,897/629/619 records); inspected pilot groups stay in training and exact duplicates cannot cross splits. Valid test sizes are 537/474/450/619 for the respective targets. We compare TF–IDF, 24-component NMF (Lee & Seung, 1999), length/style features, and frozen MiniLM, MPNet, and CLIP text, plus a separate imagery/context input channel. Neural models share 32-word chunk pooling. Logistic heads are selected by development AUROC, with preprocessing fitted on training data. AUROC measures ranking, with chance at 0.5; 95% intervals use 500 group-bootstrap resamples, conditional on fitted models. Unknown contributor overlap remains a limitation. Full protocols and additional metrics are in Appendices I and B.

Content prediction is not a complete account. Figure 2 shows strong complex-content AUROC for MiniLM (0.894), MPNet (0.891), and CLIP text (0.887), versus TF–IDF (0.877). The paired MiniLM–TF–IDF difference is 0.017, with 95% interval $[-0.013, 0.048]$, so these results do not establish a reliable ranking advantage. At threshold 0.5, neural content balanced accuracy is 0.758–0.771 versus 0.590 for TF–IDF. This comparison also depends on the decision and calibration protocol.

Narrative neural AUROC is lower for vividness (0.645–0.674), persistence (0.605–0.640), and unpleasantness (0.517–0.528); TF–IDF is competitive. This is compatible with the source prompt emphasizing visible content. Questionnaire responses may prime descriptions or share vocabulary with them. Strong content prediction is therefore a measurement check, not evidence that the full experience has been captured.

Additional reporting channels improve access to vividness. Imagery/context responses predict vividness at 0.740 AUROC. Adding these inputs raises each neural narrative model to 0.723–0.733, compared with 0.645–0.674 for text alone (Figure 3a). MPNet-plus-context reaches 0.733, a difference of -0.007 $[-0.050, 0.039]$ from context alone. Thus access to the additional channel helps relative to the text-only models; concatenation does not establish a gain over the context baseline.

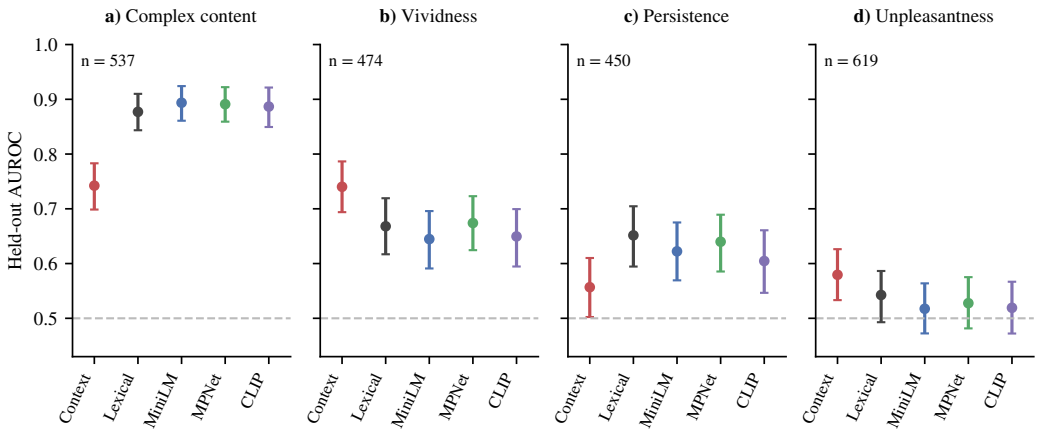


Figure 2: **Content is easier to predict than other reported qualities.** Held-out AUROC for (a) complex content, (b) vividness, (c) persistence, and (d) unpleasantness, with 95% group-bootstrap intervals. Context uses separate imagery and recording-condition responses; other models use the narrative. These are original questionnaire targets, not model-generated labels.

These associations are consistent with Ganzflicker imagery research (Reeder, 2022). For an atlas, the implication is practical: keep the contributor's vividness and duration responses alongside the narrative instead of inferring every quality from the nouns they use.

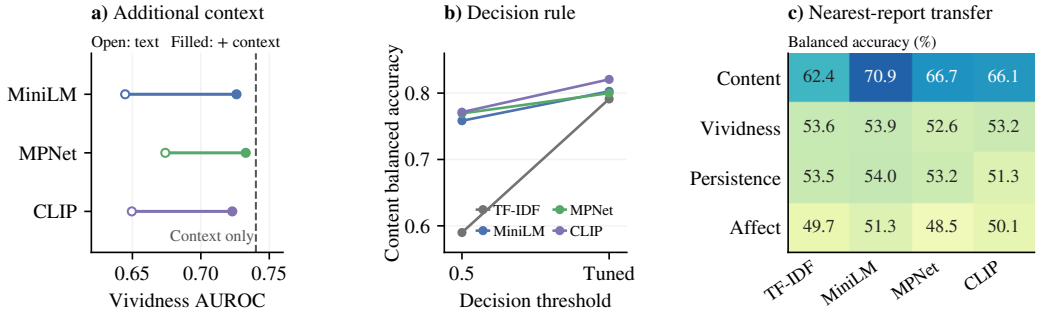


Figure 3: **Available information, decision rules and retrieval answer different questions.** (a) Adding imagery/context improves vividness prediction relative to text alone; the dashed line is context only. (b) Selecting a threshold on development data largely closes the content-classification gap between TF–IDF and neural encoders. (c) A new held-out nearest-report audit: transferring the closest training report's questionnaire label preserves broad content better than the other qualities. Values are balanced accuracy percentages (chance 50%); complete intervals are in Table 11.

The main pattern survives alternative representations and splits. Under a temporal group split, neural content AUROC remains 0.889–0.911 and affect 0.508–0.539. Removing explicit task-related words lowers lexical content AUROC from 0.877 to 0.859. A compact 24-component NMF representation retains content information (0.804) but little affect prediction (0.487). This extends the result beyond one encoder: compression and representation choice change performance without making the four reported dimensions interchangeable. Topic loadings, permutation controls and temporal results are retained in Appendices B and C and Figures 10 and 11. These checks remain within the same source population.

Decision rules explain part of the apparent encoder gap. Refitting the original heads exactly reproduces all 16 saved probability vectors. A development-selected threshold raises TF–IDF content balanced accuracy from 0.590 to 0.791; neural values become 0.800–0.821 (Figure 3b). Development-fitted sigmoid calibration lowers TF–IDF content Brier error from 0.142 to 0.114. Neither adjustment uses test labels. The changes are task-dependent: tuned thresholds lower all three neural persistence

scores. The primary AUROC result is unchanged, while the fixed-threshold comparison clearly combines representation and decision-protocol effects (Appendix M.1).

A similar real report is not an interchangeable experience. We next test the operation an atlas might use to suggest related accounts. For each target, every valid held-out query retrieves the highest-cosine training narrative from records with that target observed; query labels never enter selection. Transferring the neighbour's response gives neural balanced accuracy 0.661–0.709 for content, but only 0.526–0.539 for vividness, 0.513–0.540 for persistence and 0.485–0.513 for affect (Figure 3c). All non-content 95% group-bootstrap intervals include 0.5. TF–IDF shows the same broad pattern. This is an exploratory test of a candidate semantic-search operation, not a deployed feature or human relevance judgment. It directly supports showing separate questionnaire fields when presenting a similar account. On the same MPNet vividness queries, a development-threshold linear head improves balanced accuracy by 0.115 (paired 95% interval $[0.056, 0.172]$). Appendix N reports all paired comparisons.

5 Study 2: separate wording effects from meaning

An account's qualification can change while its nouns stay the same: "I saw a shape" and "I imagined a shape" concern similar content but different claims. We test whether a similarity readout favours a faithful paraphrase over an opposite qualification.

A counterbalanced diagnostic. The original 672 synthetic comparisons combine 24 contents, seven operators, two polarities, and two distractor styles. Each polarity has an anchor and a paraphrase. The faithful option always uses the other wording, while the distractor either resembles the query's wording or uses the paraphrase. This asymmetry can favour an opposite meaning simply because its wording overlaps more. We therefore also reverse anchor and paraphrase roles, giving 1,344 comparisons per encoder without generating new strings or embeddings. Figure 4 averages both query wordings separately for (a) a query-matched distractor and (b) alternatives written in the same style. Synthetic statements remain AI-authored; reversing roles controls this asymmetry without validating their semantics.

For frozen encoder f, the faithful-choice margin is

$$m(q, p, n) = \cos(f(q), f(p)) - \cos(f(q), f(n)), \tag{2}$$

where q is the query, p the faithful option, and n the opposite qualification. Positive margins count as correct, with half credit for numerical ties. These are text-to-text diagnostics of a readout, not an evaluation of a generator's internal encoder.

Balancing wording changes the result. Averaged across operators, accuracy with query-matched distractors versus matched alternatives is 37.2% versus 78.4% for MiniLM, 63.7% versus 89.1% for MPNet, and 20.4% versus 61.5% for CLIP text. The lexical reference changes from 14.6% to 50.0%. These are paired design sensitivities on the same templates, not gains from training a better model. Counterbalancing also prevents one query wording from determining the conclusion. In the original paraphrased condition, MiniLM's 50% motion accuracy decomposes into 100% for one polarity and 0% for the other: a persistent preference, not uniformly random performance. Appendix M.2 provides the complete polarity and wording breakdowns.

A poor similarity score does not establish missing information. Exploratory operator-specific linear probes use four content-held-out folds, training on anchor wording and evaluating paraphrases. Mean polarity accuracy is 83.9% for MiniLM, 93.5% for MPNet, and 64.0% for CLIP text; recovery varies across operators (Figure 4c; Appendix J). Probes receive supervision and solve a different task, so these are not matched improvements over cosine. MiniLM duration illustrates the distinction: matched-alternative cosine accuracy is 89.6%, and the probe recovers polarity at 97.9%; with a query-matched distractor, cosine falls to 6.25%. In contrast, CLIP motion remains difficult even for its probe (37.5%). Some failures respond to the readout, while others persist under the tested supervision. The real-report retrieval result in Study 1 gives this distinction a practical counterpart: classifier-accessible information does not automatically transfer through nearest-neighbour search.

A constructed example explains how wording and readout can interact. Let unit content vectors satisfy $u^\top u' = \rho < 1$, and encode a qualification in a last coordinate of size $\epsilon > 0$. For $q = (u, \epsilon)$,

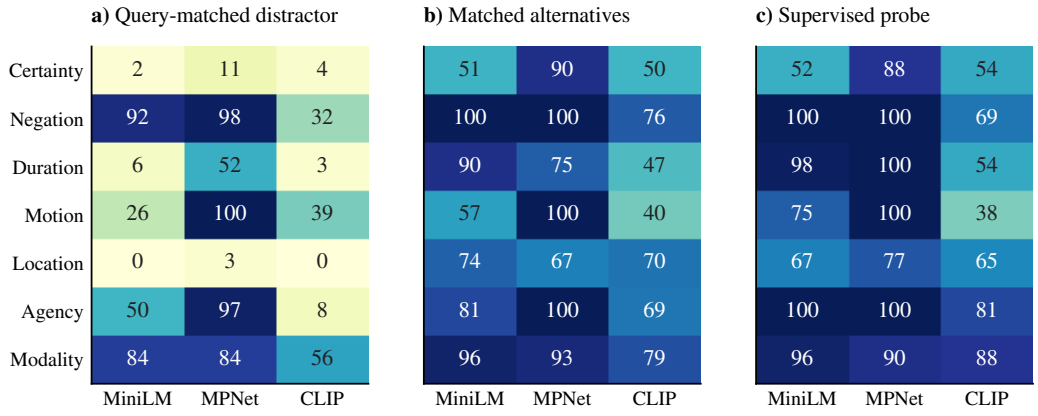


Figure 4: **Meaning depends on both comparison design and readout.** (a) Cosine choice when the opposite meaning shares the query wording. (b) Cosine choice when both alternatives use the other wording. Each cell averages both query wordings, 24 contents and two polarities (96 comparisons). (c) Original supervised polarity probes, evaluated on held-out contents and paraphrase wording (48 decisions per cell). All values are percentages. The probes use different supervision and targets; panel (c) is evidence of recoverability, not a matched improvement over (a) or (b).

$p = (u', \epsilon)$ and $n = (u, -\epsilon)$,

$$\cos(q, p) - \cos(q, n) = \frac{\rho - 1 + 2\epsilon^2}{1 + \epsilon^2}. \tag{3}$$

Cosine favours the wrong qualification if $2\epsilon^2 < 1 - \rho$, even though reading the last coordinate recovers it perfectly. If the distractor instead uses $n = (u', -\epsilon)$, the margin becomes $2\epsilon^2/(1 + \epsilon^2) > 0$. The example isolates the wording advantage by construction; it is not a fitted mechanism of our encoders. In plain language, a diagnostic can mistake closeness of wording for preservation of meaning.

6 Study 3: What do an image and its score preserve?

The first two studies motivate preserving qualifications but do not evaluate image generation. Here we compare visible-attribute coverage across the three prompt conditions, then inspect whether the metric asks and answers the right question.

Matched generation. We select 24 unique test groups with 12–70-word accounts by fixed salted text hash before annotation or generation: 74 source-linked annotations, 66 visible-detail coverage questions and three unsupported-literalization probes. The initial plan requested 144 images (two per arm); a budget amendment after 62 outputs added 22 missing-arm calls. The primary analysis uses the earliest output in every report/arm pair: 72 images, with 12 extras retained for sensitivity. At amendment time, 145 of the eventual 244 original judge ratings already existed. The selection rule excludes image-quality choices; prompts and questions remain frozen.

Output metadata identifies the backend as `gpt-image` 2.0; exact weights, seeds and internal prompt processing are unavailable. Calls are matched, not latent draws. Qwen2.5-VL-3B-Instruct (Qwen Team, 2025) judges images with fixed questions, without condition labels, prompts or expected answers. It passes 32 elementary pipeline fixtures. Full execution and budget details appear in Appendices K and G.

Coverage and uncertainty. Let y_{iak} be the judge's answer to supported question k for report i and arm a, with m_i supported questions. We compute

$$c_{ia} = \frac{1}{m_i} \sum_{k=1}^{m_i} \mathbf{1}\{y_{iak} = \text{yes}\}, \qquad \widehat{\Delta}_{S-D} = \frac{1}{24} \sum_{i=1}^{24} (c_{iS} - c_{iD}). \tag{4}$$

Uncertain/unparseable answers are not confirmed. Equal report weighting prevents longer question lists from dominating; one changed answer shifts the mean by $100/(24m_i)$ percentage points. The

primary comparison is $S - E$. Intervals use 10,000 paired report-bootstrap resamples; tests enumerate nonzero report-difference signs exactly. The unit is 24 accounts. Questions measure selected visible details; temporal qualities and emotion are outside their scope.

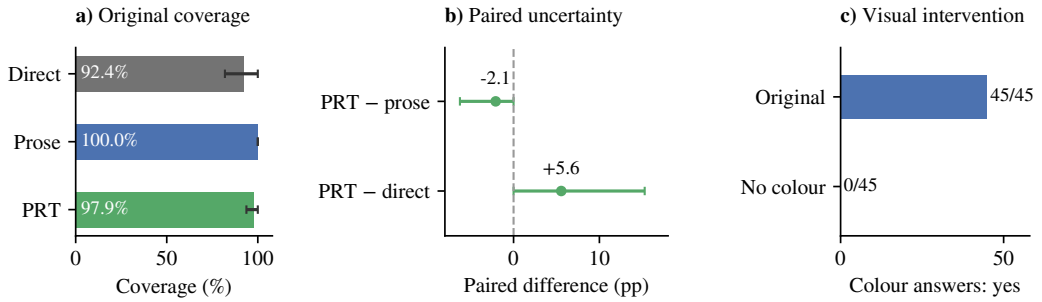


Figure 5: **What the generation comparison establishes.** (a) Original coverage with 95% report-bootstrap intervals. (b) Paired structured-PRT differences; the evidence-matched prose comparison is primary. (c) All 45 eligible colour answers change from yes to no after colour removal; rerunning the originals reproduces all yes answers. Panels (a,b) use 24 reports; panel (c) uses 15 questions from 13 reports. This intervention establishes responsiveness to removed colour, while the source/question inventory tests a separate issue: which meanings the criterion covers.

Coverage is high, but the comparison is weakly informative. Coverage is 92.4% direct, 100.0% prose and 97.9% PRT (Figure 5a,b). The PRT–prose difference is -2.1 percentage points (95% interval $[-6.2, 0.0]$, $p = 1.000$); PRT–direct is 5.6 points ($[0.0, 15.3]$, $p = 0.500$). All 207 answers are parseable and certain. The three literalization probes are negative in every arm; all-output sensitivity gives identical macro scores. Prose has no judged misses among 66 coverage questions, PRT has one and direct three. Its collapsed bootstrap interval reflects this observed ceiling, not perfect population performance. With only one and two nonzero report differences, the smallest conditional two-sided p-values are 1.0 and 0.5. Neither superiority nor equivalence is established.

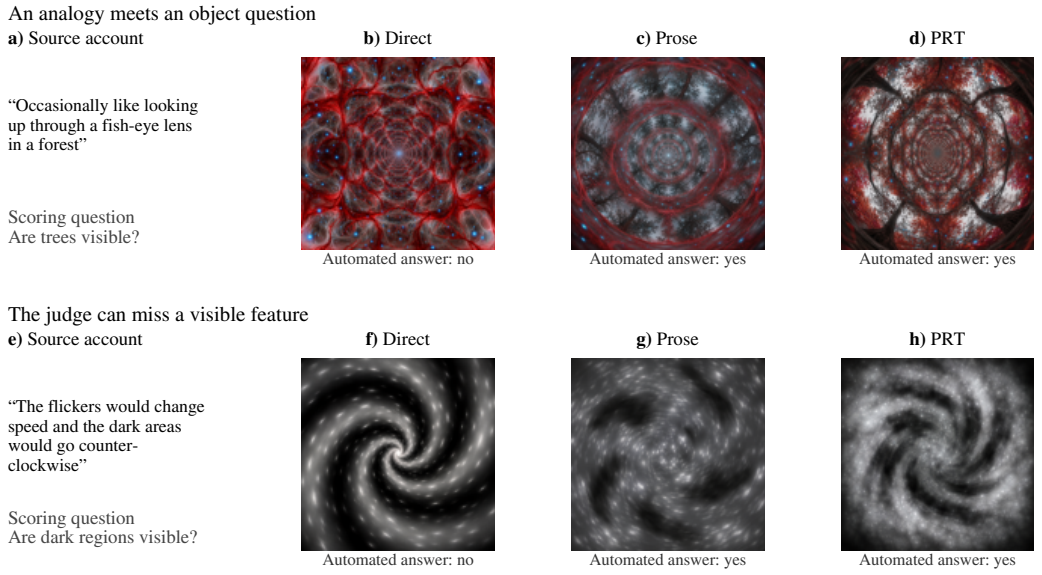


Figure 6: **Different failures require different checks.** (a–d) The forest analogy: (a) source and question; (b) direct, (c) prose and (d) PRT images. Visible trees satisfy the question without resolving the intended analogy. (e–h) The dark-region case: (e) source and question; the judge answers no for (f) direct despite visible dark regions, and yes for (g) prose and (h) PRT. These post hoc cases use the unchanged earliest outputs. Full wording is retained with the data.

Inspect every contribution to the difference. Report 3659 compares its appearance to looking through a fish-eye lens in a forest. The question requires trees, so the tree-containing prose/PRT outputs score yes and direct scores no (Figure 6). This is an ambiguous interpretation, not proof that either depiction is better. Report 5403 exposes a separate concern: the judge misses visible dark regions in the direct image.

Table 1 shows why these cases matter quantitatively. PRT and direct tie on 22 reports; the dark-region case contributes 4.17 points and the forest case 1.39 points, accounting for the entire 5.56-point difference. Against prose, PRT ties on 23 reports and scores lower on one. This exhaustive decomposition localizes the measured advantage without deleting cases or substituting new labels. A hypothetical correction of the dark-region answer alone reduces the difference to 1.39 points (Appendix M.5).

Table 1: **Every unequal report-level score.** Original confirmed coverage and each row's contribution to PRT-minus-direct (percentage points). The other 21 reports have perfect scores in all arms. These are judge outputs, not adjudicated fidelity.

Report	Questions	Direct	Prose	PRT	Contribution
1591	2	50.0%	100.0%	50.0%	+0.00 pp
3659	3	66.7%	100.0%	100.0%	+1.39 pp
5403	1	0.0%	100.0%	100.0%	+4.17 pp
Other 21 reports	—	100.0%	100.0%	100.0%	0.00 pp

Traceability and criterion coverage are separate. All 74 annotation spans occur verbatim in their sources. Nine annotations across eight reports are explicitly qualified: four uncertain, three analogical, one interpretive and one memory. None of the 69 questions directly tests communication of that status; the three literalization probes check selected objects. Five *reported*-tag spans also contain qualification cues, warranting semantic inspection rather than automatic relabelling (Appendix M.3). Both evidence arms receive interpretations also used to write scoring questions. Their comparison with direct prompting therefore measures an intervention against its own criterion. Independent source-based questions are necessary to establish greater source fidelity; the structured-versus-prose contrast isolates presentation more closely.

The judge does respond to removed visual evidence. We select all 15 chromatic-colour questions across 13 reports: 45 image–question pairs from 39 primary images. Converting images to equal RGB channels removes chromatic colour while retaining spatial structure. Every answer changes from yes to no; a contemporaneous rerun of all unmodified pairs reproduces the original yes answers (Figure 5c). Thus indiscriminate agreement does not explain this colour result. Yet a colour-responsive judge can still evaluate an incomplete criterion: none of these questions measures memory or uncertainty status. Appendix M.4 retains the exploratory protocol, all responses and the achromatic-image example (Figure 18).

7 Discussion and limitations

What the evidence makes possible. The studies suggest a concrete design for perceptual communication. Keep questionnaire responses as a separate channel because narrative models and nearest neighbours recover the four targets unevenly. Evaluate the actual retrieval rule because a qualification recoverable by a probe can lose under cosine. Keep source spans and status tags available beside illustrations because a responsive visual judge can still score only a narrow part of the account. These are complementary roles: prediction organizes a collection, retrieval proposes a comparison, and provenance lets a person inspect and dispute the interpretation.

An atlas that preserves the differences. Figure 7 shows the working record view, with the source and questionnaire responses beside evidence notes and all three prompt conditions. Readers can search 3,145 accounts, filter by vividness or duration, compare reports, and inspect the prompts and retained outputs for the 24 illustrated cases. This turns the result into a usable research artifact: similarity proposes a starting point, while separate fields preserve differences that the similarity score can overlook. A training-set example with high MPNet similarity but different vividness and persistence appears in Appendix Figure 12; the held-out retrieval audit establishes that the concern is

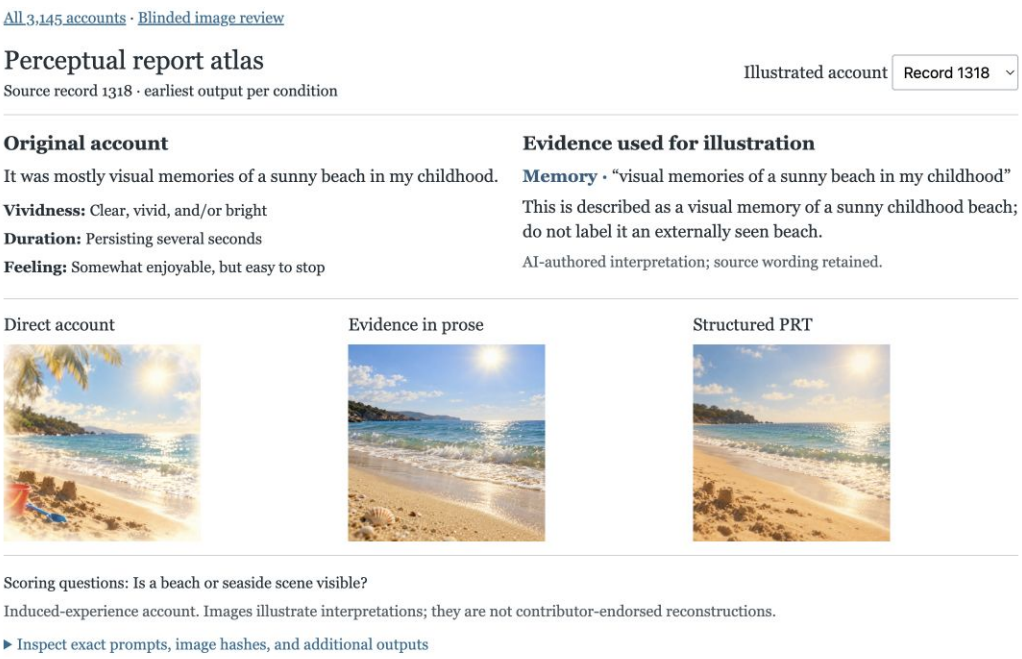


Figure 7: **The working atlas: inspect an illustration against its source.** Screenshot of the offline record view for account 1318, selected to illustrate a memory qualification. The account and questionnaire responses appear beside the evidence record; below are the earliest outputs from all three conditions. The memory status stays explicit even though each image depicts a beach. The expandable control exposes exact prompts, image hashes, and retained repeats. This is an interface demonstration, not a reconstruction endorsed by the contributor.

broader than that selected pair. The offline view is `atlas/browser/illustration.html`; a separate blinded review interface is supplied for future ratings.

Limits of the evidence. The sample is single-source and self-selected, with heterogeneous viewing conditions and unknown contributor overlap. It supplies neither clinical validation nor endorsed image targets. Synthetic diagnostics cover seven AI-authored template families; probes, context models and retrieval extensions are exploratory. The image pilot uses short selected accounts, one backend and one judge, with AI-authored interpretations and uncontrolled hidden processing or backend drift. Its criterion covers selected visible details. Contributor endorsement, reader understanding and clinical usefulness remain unmeasured.

The next outcome is reader understanding of the complete presentation: account, image, and explanatory context. A useful test would compare account-only, account-plus-image, and matched explanatory-caption conditions, measuring supported and unsupported inferences. Captions matter: a beach image may legitimately illustrate a memory, and an analogy can legitimately guide a drawing. The failure is an unjustified interpretation communicated to a reader, not every detail absent from the source. Independent source adjudication and contributor feedback should determine acceptable interpretations before scoring methods against them.

8 Conclusion

PRT connects perceptual accounts to representations and illustrations while keeping their evidence inspectable. Prediction, retrieval, controlled readouts and image interventions show why resemblance alone cannot establish preserved experience. The resulting atlas supports comparison without making different reported qualities interchangeable. Its next test is whether the complete presentation improves understanding without adding unjustified inferences.

A bridge between experiences begins with something another person can recognize, while leaving room for what only the speaker can clarify. Its value lies in making an unfamiliar experience discussable, with the person describing it still able to say what the image gets wrong.

AI ASSISTANCE DISCLOSURE

Generative AI assisted with conceptual development, experimental design and implementation, data processing and annotation, synthetic data and image generation, analysis, and manuscript preparation.

REPRODUCIBILITY STATEMENT

The accompanying code includes source checksums, selection and split rules, model revisions, exact generation prompts, all outputs, scoring questions and responses, analysis scripts, and figure generation. Protocol and request hashes were fixed before image generation. The source data are available at `https://osf.io/6dvh9/`. No new participant data were collected. Appendix G gives execution details.

REFERENCES

Romy Beauté, David J Schwartzman, Guillaume Dumas, Jennifer Crook, Fiona Macpherson, Adam B Barrett, and Anil K Seth. Mapping of subjective accounts into interpreted clusters (mosaic): topic modelling and llm applied to stroboscopic phenomenology. *Neuroscience of Consciousness*, 2026 (1):niag008, 2026.

Chuer Chen, Zichen Wang, Yi He, Zhengxi Yu, and Nan Cao. Blending concepts: Benchmarking visual metaphor generation in text-to-image models. *arXiv preprint arXiv:2609.02502*, 2026.

Ana Chkhaidze, Reshanne R Reeder, Connor Gag, Anastasia Kiyonaga, and Seana Coulson. From dots to faces: individual differences in visual imagery capacity predict the content of ganzflicker-induced hallucinations. *Neuroscience of Consciousness*, 2026(1):niag016, 2026.

Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. Towards understanding and quantifying uncertainty for text-to-image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8062–8072. IEEE, 2025.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. Clip behaves like a bag-of-words model cross-modally but not uni-modally. In *International Conference on Learning Representations*, volume 2026, pp. 96106–96127, 2026.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

Youngsun Lim, Hojun Choi, and Hyunjung Shim. Evaluating image hallucination in text-to-image generation with question-answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26290–26298, 2025.

Chen-Yi Lu, Yueh-Shao Chen, and Somali Chaterji. What clip knows but cannot say: Recovering negation from frozen intermediate features. *arXiv preprint arXiv:2607.23271*, 2026.

Katie Melvin, Jon Crossley, and John Cromby. The feeling, embodiment and emotion of hallucinations in first episode psychosis: A prospective phenomenological visual-ecological study using novel multimodal unusual sensory experience (muse) maps. *EClinicalMedicine*, 41, 2021.

Rosiana Natalie, Wenqian Xu, Ruei-Che Chang, Rada Mihalcea, and Anhong Guo. Not there yet: Evaluating vision language models in simulating the visual perception of people with low vision. *arXiv preprint arXiv:2508.10972*, 2025.

Vincent Quantmeyer, Pablo Mosteiro, and Albert Gatt. How and where does clip process negation? In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pp. 59–72, 2024.

Qwen Team. Qwen2.5-VL, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

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.

Reshanne R Reeder. Ganzflicker reveals the complex relationship between visual mental imagery and pseudo-hallucinatory experiences: A replication and expansion. *Collabra: Psychology*, 8(1): 36318, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Candace Ross, Melissa Hall, Adriana Romero Soriano, and Adina Williams. What makes a good metric? evaluating automatic metrics for text-to-image consistency. *arXiv preprint arXiv:2412.13989*, 2024.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Luecheng Wang, Shintaro Ozaki, Hidetaka Kamigaito, Katsuhiko Hayashi, Jingun Kwon, Manabu Okumura, and Taro Watanabe. Simile understanding in text-to-image models: An evaluation framework. *arXiv preprint arXiv:2608.04750*, 2026.

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.

A Data provenance and label mappings

The input is `Ganzflicker_Experience6k.csv`, OSF file `614a20b147d8290018671c0b`, version 1. Its SHA-256 is `0a716e983ea4b49d0df83af37c6e33c229dd6b5bb9c4308b78bffc28322ceaca`. The original README is version 2 and also matches its OSF checksum. We retain original bytes and decode Windows-1252; UTF-8 decoding fails. All 6,665 logical records have 35 fields. Seven unnamed columns are empty. There are no entirely blank rows, repeated headers, or exact duplicate whole rows. These observations do not resolve the discrepancy with the documented count of 6,664.

The sequential eligibility exclusions are 90 records without explicit consent, 940 without an age in the specified range, and 2,490 outside the 5–400-word description range, leaving 3,145. These exclude neither all unusual answers nor all nonstandard viewing contexts. We do not imply reproduction of the original paper's 1,810-record final sample.

Target	Train	Development	Test (positive)
Content	1650	550	537 (144)
Vividness	1479	483	474 (180)
Persistence	1406	458	450 (281)
Affect	1897	629	619 (234)

Table 2: Valid task-specific record counts in the primary split. Missing targets are not imputed.

Content. The allowed selections are distortions, simple shapes/patterns, colors other than red and black, complex objects, and complex environments, using the full exact strings in the source. Positive means one of the last two options is selected. Negative means a valid other selection or exactly "No." Residual free text and combinations containing "No" plus other options are excluded rather than reinterpreted. **Vividness.** The ordered source options are weak/faint/insubstantial; clear but not vivid; clear and moderately vivid; clear/vivid/bright; and very vivid/almost real/popping out. Only an exact single option is valid. The last three are positive. **Persistence.** The ordered options are brief moment/flash; 1–2 seconds; persisting several seconds; and constant, morphing from one image to the next. Only an exact single option is valid. The last two are positive. This is a coarse threshold, not an assertion that constant morphing and several seconds are identical experiences. **Affect.** The two unpleasant options are positive. Neutral, somewhat enjoyable, and enjoyable are negative. Nonstandard strings are excluded. The target concerns the session's emotional rating, which may include aspects not described in the visual account.

Normalization for split grouping lowercases and collapses whitespace; it does not remove punctuation or identify paraphrases. The SHA-256 of `prt-study-v1:` concatenated with the normalized-text hash determines the primary split. Twelve inspected pilot groups are forced to training. For temporal sensitivity, groups are sorted by latest source-row occurrence; the first 60% train, next 20% develop, and last 20% test. Pilot groups are excluded there. This avoids forcing late inspected examples into temporal training. Source timestamps are monotone but may reflect collection and questionnaire changes.

B Additional metrics and sensitivity

The style baseline uses log word count, log character count, counts of commas, periods, question marks, and digits, and the type/token ratio. Context features contain imagery ratings and five recording-condition indicators; missing numeric values use the training median. Dense inputs are standardized from training data. TF–IDF uses unigrams/bigrams, minimum document frequency 3, maximum document frequency 0.98, sublinear term frequency, and at most 12,000 features. The masked variant removes a declared list of word stems associated with the target descriptions before fitting its own vocabulary. It does not remove all semantic synonyms.

The ten exploratory primary-split permutation runs independently shuffle training and development labels, select C on shuffled development labels, and evaluate against unchanged test labels. Their AUROC ranges are 0.400–0.599 for content, 0.448–0.540 for vividness, 0.412–0.568 for persistence,

Task	Model	AUROC (95% interval)	Balanced acc.	Brier
Content	TF–IDF	0.877 [0.843, 0.910]	0.590	0.142
Content	Imagery + context	0.742 [0.699, 0.783]	0.516	0.173
Content	MiniLM	0.894 [0.861, 0.924]	0.758	0.108
Content	MPNet	0.891 [0.859, 0.922]	0.769	0.112
Content	CLIP text	0.887 [0.849, 0.921]	0.771	0.107
Vividness	TF–IDF	0.668 [0.617, 0.719]	0.500	0.235
Vividness	Imagery + context	0.740 [0.694, 0.787]	0.645	0.197
Vividness	MiniLM	0.645 [0.591, 0.696]	0.573	0.229
Vividness	MPNet	0.674 [0.624, 0.723]	0.611	0.225
Vividness	CLIP text	0.650 [0.595, 0.699]	0.580	0.231
Persistence	TF–IDF	0.652 [0.595, 0.705]	0.500	0.231
Persistence	Imagery + context	0.557 [0.502, 0.610]	0.521	0.235
Persistence	MiniLM	0.622 [0.569, 0.675]	0.605	0.236
Persistence	MPNet	0.640 [0.585, 0.689]	0.601	0.233
Persistence	CLIP text	0.605 [0.547, 0.661]	0.602	0.240
Affect	TF–IDF	0.543 [0.493, 0.586]	0.500	0.235
Affect	Imagery + context	0.579 [0.533, 0.626]	0.522	0.232
Affect	MiniLM	0.517 [0.472, 0.564]	0.515	0.254
Affect	MPNet	0.528 [0.482, 0.575]	0.523	0.257
Affect	CLIP text	0.519 [0.472, 0.567]	0.515	0.256

Table 3: Primary held-out metrics. Intervals resample known normalized-text groups and are conditional on fitted models. Balanced accuracy uses the unadjusted probability threshold 0.5.

Split	Representation	Content	Vividness	Persistence	Affect
Primary	TF–IDF	0.877	0.668	0.652	0.543
Primary	Imagery + context	0.742	0.740	0.557	0.579
Primary	MiniLM	0.894	0.645	0.622	0.517
Primary	MPNet	0.891	0.674	0.640	0.528
Primary	CLIP text	0.887	0.650	0.605	0.519
Temporal	TF–IDF	0.878	0.639	0.616	0.536
Temporal	Imagery + context	0.711	0.735	0.589	0.595
Temporal	MiniLM	0.911	0.665	0.653	0.516
Temporal	MPNet	0.889	0.682	0.623	0.508
Temporal	CLIP text	0.903	0.663	0.621	0.539

Table 4: AUROC under the primary and temporal group splits. These are two evaluations of the same source, not independent external replication.

and 0.469–0.533 for affect. These controls show finite-sample variation rather than certifying the absence of every form of leakage.

C Topic representation

NMF uses 24 components, NNDSVDa initialization, random seed 20260912, and at most 400 iterations, fitted to training TF–IDF. The dimension is fixed rather than selected for an attractive visualization. Example high-loading terms include "shapes, geometric shapes, geometric"; "blue, yellow, green"; and "light, balls, balls of light." Another component begins with "but, were, they, there," demonstrating that a factor can capture reporting language rather than a perceptual category. Component labels are AI interpretations, not validated human coding. All 24 term lists are saved with the results.

D Synthetic contrasts and probe protocol

Table 5 lists the anchor pairs. Each has an independently written faithful paraphrase; the full strings are included in the code. For example, certainty paraphrases are "I am certain that x appeared" and "I am uncertain whether x appeared." Motion paraphrases are "The x kept changing position" and

"The x stayed completely still." The wrong option is either the other anchor or its paraphrase. The second form removes some asymmetric lexical-overlap advantage; it does not create a fully balanced natural-language benchmark.

Operator	Anchor 0	Anchor 1
Certainty	I definitely saw x.	I might have seen x, but I am not sure.
Negation	I saw x.	I did not see x.
Duration	I saw x continuously.	I saw x for just an instant.
Motion	The x were moving.	The x were stationary.
Location	I saw x in the center of my view.	I saw x at the edges of my view.
Agency	I could make x appear at will.	I could not make x appear at will.
Seeing/imagining	I saw x on the screen.	I imagined x in my mind.

Table 5: AI-authored template anchors. Contents are plural noun phrases. These statements are synthetic and not attributed to participants.

The 24 contents include colors and simple forms as well as faces, animals, buildings, trees, flowers, windows, ribbons, tunnels, clouds, rings, shadows, and human figures. We avoid inherently moving nouns in stationary templates. A content-control distractor preserves the query form but substitutes the content 12 positions away in the fixed list. Thus it changes both object identity and sometimes appearance; it is not matched to an operator change in edit distance or perceptual magnitude.

Faithful-choice intervals resample the 24 content families 2,000 times, keeping their polarity trials together. They quantify variation across the selected contents, not across new template families. We emphasize per-operator results rather than treating every template-generated trial as an independent observation.

For polarity probes, four consecutive six-content blocks rotate through test and development roles; the other two blocks train. Training uses anchors, whereas development and test use paraphrases. Each operator has its own standardized logistic probe with the same four C values. This probes generalization across content and wording, but not to an unseen operator or human-authored paraphrase. Its accuracy is not directly comparable to unsupervised two-candidate retrieval, which uses different inputs and supervision.

E Development demonstrations

A preliminary compiler was run on three purposively selected development records from the existing pilot, covering spatial appearance, temporal change, and uncertainty (Figure 8). These induced visual experiences provide real accounts for demonstrating the workflow, without assigning a psychiatric or neurological diagnosis. Full source text, annotations, compiled prompts, and outputs are retained in the atlas demonstration. This development compiler used a stricter rendering policy, including abstention for an account without affirmative visual attributes. It is archived separately; the controlled comparison uses the policy in Section 2 and generates an image for every condition.

The process reveals an annotation failure as well as a possible use. For row 1568, the initial pilot record contained colors but omitted "strong lines." The first image therefore showed colored shapes. Adding the missing lines still produced unsupported circles and triangles; preserving the source's "laser lightshow" description yielded the displayed illustration. All three attempts are archived. Row 691 used one generation call; row 4007 used none. This is an iterative demonstration, not a first-attempt success rate or a controlled comparison of generators.

The resulting record lets a reader inspect a visual interpretation alongside its source and the temporal or uncertain aspects it cannot depict. Exact layout and brightness remain rendering choices. We have not yet tested whether this record improves another person's understanding, or whether contributors prefer it to their account alone. That communication outcome is the relevant next test, rather than whether a model recognizes the intended objects.

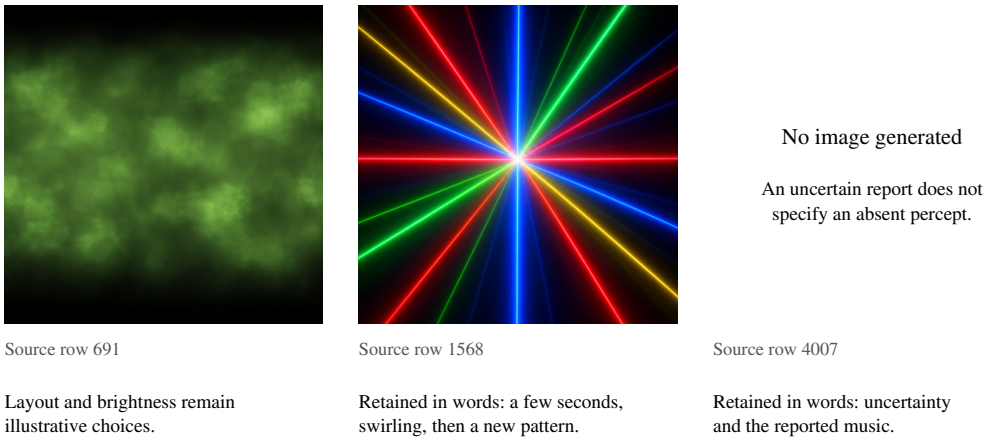


Figure 8: **From reported attributes to a communication record.** (a) Source row 691 reports diffuse greenish patches and dark upper/lower boundaries. (b) Row 1568 describes a colored laser-like pattern, lasting seconds and changing; the still image illustrates appearance while the text retains time course. (c) Row 4007 expresses uncertainty about green coloration and shapes, so PRT retains the account without generating a definitive visual depiction. Descriptions above the outputs are condensed interpretations, not verbatim quotes. Images are generated development examples, not participant-endorsed targets.

F Atlas schema and evidence boundaries

Atlas implementation and clinical translation. An experience atlas should let a reader find accounts with a particular feature, inspect their source, and compare qualities that a single diagnosis or similarity score would hide. We implement a local descriptive prototype containing the 3,145 eligible accounts, with text search, vividness and duration filters, and side-by-side comparison of original content, vividness, duration, and feeling responses. It contains induced experiences, not a clinical cohort. We use *atlas* for this organized collection; the current prototype establishes neither a taxonomy of disorders nor a validated map of subjective experience.

An experience atlas should be useful for communication, not merely recognizable to a model. The first design implication is to preserve separate fields for visual content, time course, uncertainty, context, and interpretation, with evidence spans and unknown values. A description of uncertain green coloration should not automatically become a confidently green scene. A narrative about visual static should not automatically acquire an auditory component because it uses a sound-related analogy.

The second implication is to evaluate the operation actually used. If an atlas retrieves neighboring reports by cosine similarity, a classifier's success is insufficient validation. If a tool generates illustrations, text-encoder accuracy is insufficient validation. A future reconstruction study should ask contributors what a depiction gets right, gets wrong, or cannot represent, then test whether it improves another person's understanding of the reported difficulties. The current experiments establish neither participant endorsement nor such communication benefit.

For clinical translation, the next decisive test is whether contributors endorse an illustration and whether another reader understands the reported difficulty more accurately from the account plus illustration than from the account alone. This would assess communication rather than presumed access to a private percept.

The third implication is to preserve the distinction between accounts, questionnaire responses, annotations, and generated media. PRT should support communication while keeping each interpretation traceable to its evidence.

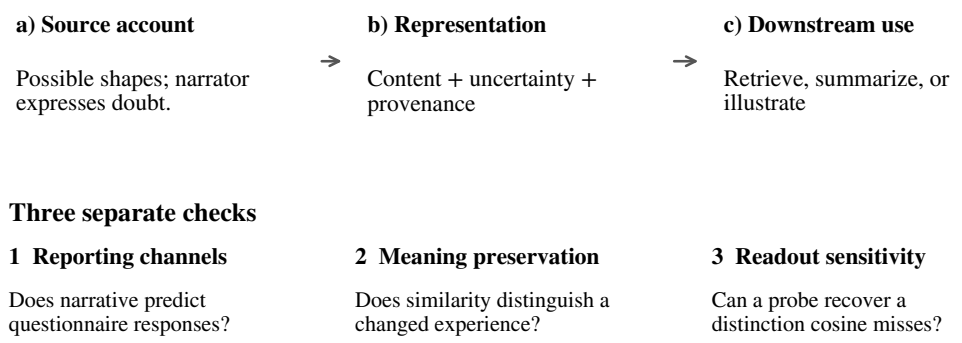


Figure 9: **Three checks before an illustration is trusted.** (a) Source account, (b) representation, and (c) downstream use, with separate checks for reporting channels, similarity ordering, and readout sensitivity. It is a design aid, not a validated model of subjective experience.

The development pilot contains 12 deterministically selected original accounts and 47 AI-authored annotations. Each annotation includes its interpretation, reported/uncertain/negated status, and an exact source-text span. Human annotations are a separate empty field. No participant has endorsed these annotations or the generated illustrations. The pilot is not a clinical corpus or a gold-standard test set.

Atlas records should distinguish a source account, optional diagnostic context explicitly supplied by a contributor, researcher annotations, model annotations, and generated media. They should retain reuse terms and the provenance of any pairing. Unknown attributes and aspects that cannot be depicted should remain visible. The browser reproduces source responses rather than replacing them with binary study labels or model annotations. It omits demographic fields and does not infer diagnoses. The present empirical analysis uses original questionnaire answers, not pilot annotations, and does not train a report-to-image generator. Figure 1 is an author-specified illustration of the communication goal. Figure 8 uses source-linked pilot records; generation prompts, annotation edits, and all image attempts are archived.

The MUSE publication is methodological background only: its data-sharing statement says participants did not agree to dataset sharing (Melvin et al., 2021). A public image-recreation deposit was identified but its dataset reuse terms and record pairing were not established in this audit. The current MOSAIC repository supplies analysis code but no raw accounts in its checked data folders. The source audit records these distinctions without presenting all readable publications as available datasets.

G Execution and reproducibility

Experiments run locally on Apple silicon with 64 GB memory, Python 3.11.13, PyTorch 2.8.0, Transformers 4.56.1, Sentence Transformers 5.1.0, and scikit-learn 1.7.1. Neural inference uses the MPS backend. Models remain frozen; logistic probes are fitted locally. The numerical encoder experiments use local inference. The development demonstration sends selected source-linked visual excerpts to the image-generation backend; the controlled comparison sends the full selected accounts with condition-specific instructions.

Model identifiers are `sentence-transformers/all-MiniLM-L6-v2`, `sentence-transformers/all-mpnet-base-v2`, and `openai/clip-vit-base-patch32`. Output dimensions are 384, 768, and 512. Checkpoint cards, source metadata, derived-file checksums, and execution logs accompany the local research package.

The execution order is source retrieval, eligibility/target preparation, frozen encoding, primary and temporal prediction, synthetic contrasts, exploratory follow-ups, and figure/table generation.

Source files remain unchanged. A separate protocol distinguishes the initial plan from exploratory additions. Tests check target mappings, split separation, embedding alignment, stored predictions, source checksums, and evidence provenance. They establish computational consistency, not human validation of the conceptual framework. Reproduction instructions and limitations accompany the code.

H CONTROLLED GENERATION AND SCORING DETAILS

The fixed protocol, 24 source records, 74 annotations, 69 questions, and 144 shuffled requests are stored separately. The original 144-call plan was amended for budget reasons after 62 outputs. Twenty-two additional calls completed the 24 three-condition groups; the earliest output per condition supplies the 72-image primary analysis. Both evidence conditions include each annotation's status, exact source span, and interpretation. Prose and structured prompts reproduce the same evidence; tests verify byte-for-byte agreement between the active compiler and the frozen requests. The original account remains verbatim in all conditions. Source-span checks reject unmatched evidence, but cannot validate an interpretation.

For illustration scoring, Qwen2.5-VL-3B-Instruct uses revision `66285546d2b821cf421d4f5eb2576359d3770cd3`, half precision on MPS, eager attention, deterministic decoding, and a maximum of 12 new tokens. Image processing uses the installed fast processor, a minimum of 256 and maximum of 512 visual tokens' worth of pixels (28^2 pixels per token). The exact prompt asks for YES, NO, or UNCERTAIN based only on visible content; neither condition labels nor expected answers enter the judge. Questions are shuffled with a fixed seed. Full raw responses are retained, including parsing failures.

Confirmed coverage weights reports equally after averaging their supported-attribute questions in the earliest output per condition. An upper bound permits uncertain or unparseable responses to count as present; this is retained with the results rather than silently recoding them as absence. The three unsupported-literalization questions concern an actual coffee vessel, train hardware, and anatomically depicted blood vessels. They are not representative of every unsupported object a generator could add.

I PREDICTION PROTOCOL AND SUPPORTING ANALYSES

I.1 SOURCE, SELECTION, AND TARGETS

We use the public Ganzflicker deposit associated with Reeder (2022). Appendix A documents source checksums, eligibility rules, and the unresolved difference between 6,665 parsed records and the source's documented 6,664.

We retain records with explicit consent, a numeric reported age from 18 to 100, and a visual description containing 5–400 whitespace-delimited words. This leaves 3,145 records and 3,134 distinct normalized texts. These rules define our analysis sample; they do not reproduce the original study's observation-condition exclusions. The data include heterogeneous viewing conditions, self-reported attributes, and no validated stable person identifier. We therefore use "records," not a claim that all retained rows identify distinct people.

The questionnaire provides four targets. *Content* indicates selection of complex objects or environments versus valid selections without these options. *Vividness* indicates an exact single answer at least "clear and moderately vivid." *Persistence* indicates an exact single answer of several seconds or constant/morphing, versus a flash or 1–2 seconds. *Affect* indicates an unpleasant rating versus neutral or enjoyable. Missing, mixed, contradictory, or nonstandard answers are excluded from the affected task only. This deliberately avoids converting ambiguous multi-selections into a single severity score. Appendix A specifies the mappings and counts.

I.2 SPLITS, MODELS, AND METRICS

Normalized-text hashes assign groups to training, development, and test in approximately 60/20/20 proportions: 1,897, 629, and 619 records. Exact text duplicates cannot cross splits. The previously

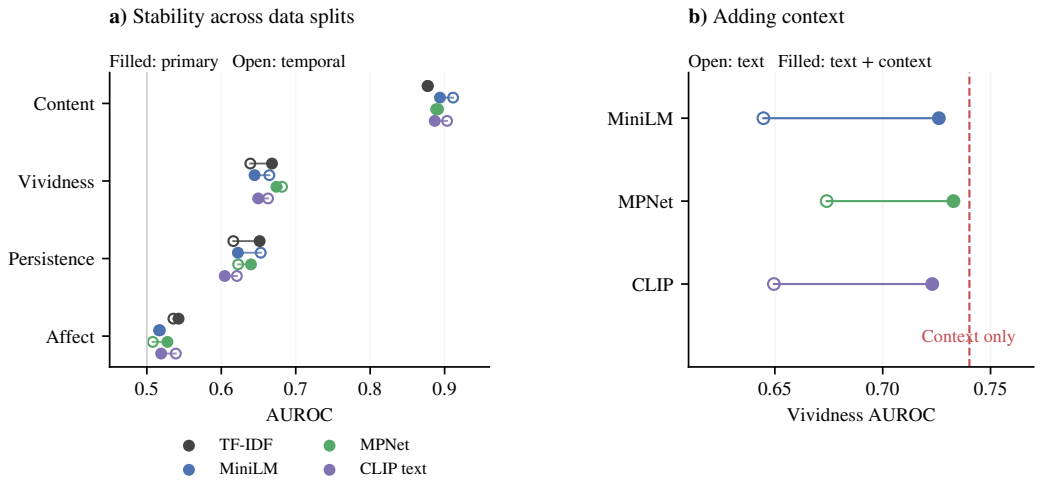


Figure 10: **The result depends on the question and available inputs.** (a) The broad content-versus-other-dimensions pattern persists under a time-ordered group split, although rankings change. (b) Exploratory addition of imagery/context inputs improves vividness prediction relative to narrative embeddings alone; it does not outperform the context-only point estimate. Intervals and paired comparisons are supplied with the results.

inspected pilot groups are forced to training. Different surface forms from the same unidentifiable contributor could still cross splits; this is a limitation of the source. The primary test sets contain 537, 474, 450, and 619 valid records for the four targets respectively.

We compare a prevalence baseline, seven length/style features, seven imagery/context covariates, word TF–IDF, a 24-component nonnegative factorization of training TF–IDF (Lee & Seung, 1999), and frozen MiniLM, MPNet, and CLIP text encoders. The context covariates contain visual and auditory imagery ratings and indicators for darkness, computer use, white noise, and viewing duration. Neither the target questionnaire answer nor diagnosis enters a narrative representation. The context model addresses a different input channel and is not advertised as a text-only competitor.

For comparable narrative coverage, all neural models encode the same non-overlapping 32-word chunks. We normalize chunk embeddings, average them equally, and normalize the result. Two of 5,611 chunks exceed CLIP's 77-token limit and are truncated; none exceed the sentence encoders' limits. The operator diagnostic uses short sentences and does not involve multi-chunk pooling. Exact model revisions and software are recorded in Appendix G.

Each feature representation uses a logistic head with $C \in \{0.01, 0.1, 1, 10\}$ selected by development AUROC. Transformations are fitted on training data only; the chosen model is not refitted on the test set. We report AUROC (how well scores rank positive above negative responses; chance is 0.5), balanced accuracy at threshold 0.5, and Brier score (probability error; lower is better). Confidence intervals resample normalized-text groups 500 times, conditional on the fitted model. Paired differences use 1,000 resamples. These intervals do not account for unknown contributor dependence or every source of model-selection uncertainty.

The analysis plan was written before fitting, after the schema pilot. It is an internal prospective plan, not a public preregistration. Chronological sensitivity, lexical masking, and label permutation check specific alternative explanations. Combined-input models and polarity probes were added after viewing initial results and are explicitly exploratory.

I.3 Full primary comparison and context sensitivity

Additional input matters more than a blanket encoder upgrade. The imagery/context model predicts vividness at 0.740 AUROC. In the exploratory concatenation experiment, narrative-plus-context AUROC is 0.733 for MPNet and 0.723 for CLIP, while MiniLM also improves over its narrative-only version. MPNet-plus-context differs from context alone by -0.007 $[-0.050, 0.039]$. Thus adding context helps relative to the corresponding narrative model, but this experiment does not

Table 6: **Primary held-out AUROC.** All models use the same task-specific records. The context row has additional inputs. Masked TF–IDF removes a prespecified set of explicit content, vividness, duration, and affect words. Full intervals and additional metrics appear in Appendix B.

Representation	Content	Vividness	Persistence	Affect
Prevalence	0.500	0.500	0.500	0.500
Length/style	0.651	0.591	0.616	0.524
Imagery + context	0.742	0.740	0.557	0.579
TF–IDF	0.877	0.668	0.652	0.543
NMF (24 topics)	0.804	0.599	0.621	0.487
MiniLM	0.894	0.645	0.622	0.517
MPNet	0.891	0.674	0.640	0.528
CLIP text	0.887	0.650	0.605	0.519
Masked TF–IDF	0.859	0.651	0.636	0.545

establish an improvement over context alone. Figure 10 makes both comparisons explicit. Known associations between imagery and induced perceptual vividness (Reeder, 2022) provide a plausible explanation; this is not a novel psychiatric predictor.

Sensitivity checks limit stronger interpretations. A temporal group split retains high content prediction (neural AUROC 0.889–0.911) and weak affect prediction (0.508–0.539). Removing explicit task-related words reduces lexical content AUROC from 0.877 to 0.859. Ten label-permutation runs show substantial finite-sample variation around chance. These within-source checks address specific alternatives, not external replication or causal identification; Appendix B gives details.

Topic compression retains content better than affect. A 24-component NMF representation reaches 0.804 content AUROC but 0.487 for affect. Components include lexical groups around geometric shapes, color, light, and the screen, as well as a component dominated by common reporting words. Figure 11 compares the topic representation with lexical and style features; Appendix C gives representative loadings. These components are interpretable descriptions of word co-occurrence, not established perceptual primitives or biological subtypes. A visually attractive map of them would not establish those stronger claims.

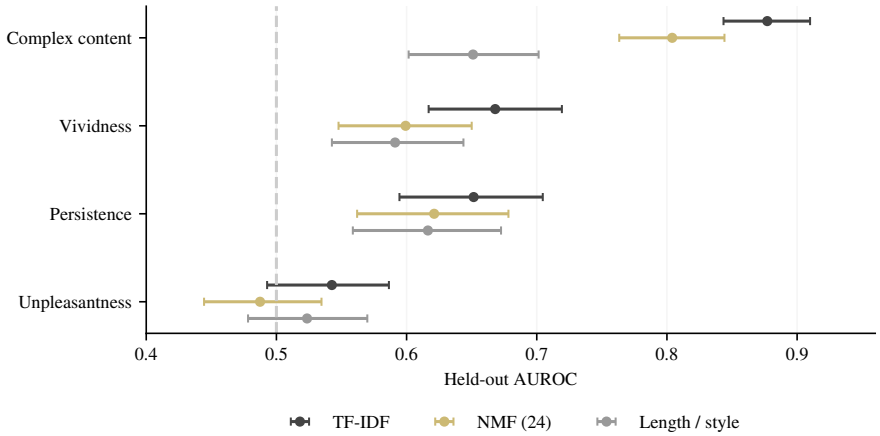


Figure 11: **A compact lexical decomposition.** Predictive performance of 24 NMF components versus full TF–IDF and style features. The representation retains some content information but does not establish a universal atlas of experiences.

J READOUT PROBES AND THEIR INTERPRETATION

J.1 EXPLORATORY PROBES RECOVER SOME DISTINCTIONS

We next fit a separate binary polarity probe for each operator using frozen representations. A probe is a simple trained classifier that reads a distinction from otherwise unchanged encoder outputs. In four folds, 12 contents provide anchor-wording training examples, six provide paraphrase development examples, and six provide paraphrase test examples. Each content appears in the test set once, producing 48 held-out decisions per operator. The operator family is known; this is not transfer to a new semantic operation.

Figure 16 shows recoverability that varies by operator and model. Averaged over the seven operators, polarity accuracy is 83.9% for MiniLM, 93.5% for MPNet, and 64.0% for CLIP. Recovery is not uniform: CLIP's motion probe remains below chance at 37.5%, while its agency probe reaches 81.3%. Thus a linear probe is a diagnostic tool, not a universal correction. Wording generalization, small content families, and operator-specific supervision constrain the result.

These observations support a careful conclusion: some similarities are poor proxies for distinctions that another readout can access. They do not show that a generator will preserve those distinctions, that a probe captures a neural mechanism, or that all failures can be fixed with a linear layer. Existing work already distinguishes encoded information from alignment and readout behavior (Koishigarina et al., 2026; Lu et al., 2026); our analysis brings that distinction into an experience-report audit.

Why the distinction matters. Similarity and recoverability answer different practical questions. A nearest-neighbor search can return an account with the wrong qualification even when a trained classifier can recover that qualification from the same representation. For example, $z = (u, \epsilon a)$ can retain an attribute a in a small coordinate while cosine is dominated by content u. This explains a possible discrepancy, not a proven mechanism of our encoders. A communication atlas should evaluate its actual retrieval or generation operation rather than borrow reassurance from another task's score.

K ILLUSTRATION PROTOCOL AND COMPLETE NUMERICAL RESULTS

K.1 REPORTS, ANNOTATIONS, AND GENERATION

The underlying source contains visual descriptions and separate questionnaire responses. We retain adult records with explicit consent and descriptions of 5–400 words, giving 3,145 records and 3,134 unique normalized texts. Normalized-text groups are assigned to training, development, and test; the earlier 12-record development pilot belongs to training. Appendix A gives eligibility, target mappings, and source provenance.

For the illustration study, we select 24 unique test groups with descriptions of 12–70 words, ordered by a fixed salted text hash. Selection precedes annotation and image generation. There are 74 source-linked annotations and 69 scoring questions. Of these, 66 concern supported visible details; three separately check specific unsupported literalizations. The latter are deliberately narrow and do not constitute a comprehensive hallucination measure.

The initial frozen plan called for two images in each condition (144 calls). After 62 outputs, generation was paused for budget reasons. A documented amendment authorized exactly 22 missing-condition calls, completing at least one image for every report and condition. The primary analysis uses the earliest recorded output for each report–condition pair: 72 images from all 24 original reports. Selection never uses output quality. The 12 additional outputs are retained for a sensitivity analysis. At amendment time, 145 of the 244 eventual judge ratings already existed; judging ran during generation. Prompts, evidence, and scoring questions remain unchanged; this is a budget-amended pilot, not the original repeated-output design. Earlier iterative development cases are excluded (Appendix E).

The backend is the built-in image-generation tool; embedded output metadata identifies `gpt-image`, version 2.0. We save every returned image, prompt hash, source path, and available model metadata. The interface does not expose an exact weights revision, random seed, or internal prompt processing. Our design therefore matches the number of generation calls, not identical latent draws or compute.

Table 7: **Controlled illustration results.** Coverage averages supported-attribute judgments within report before averaging 24 reports. The last column counts confirmed unsupported literalizations over three images per condition, from only three specific source accounts. Lower is better for that column; it is not a general hallucination rate.

Prompt	Coverage [95% interval]	Literalizations
Original report	92.4% [81.9, 100.0]	0/3
Evidence in prose	100.0% [100.0, 100.0]	0/3
PRT: structured evidence	97.9% [93.8, 100.0]	0/3

The source prompts request one square image and permit reported symbols while excluding added explanatory captions.

K.2 Measurements and uncertainty

A separate local Qwen2.5-VL-3B-Instruct model (Qwen Team, 2025) answers fixed questions about each image, without condition labels, original generation prompts, or expected answers. Questions concern visible properties such as a red ball, a grid, or an upper-left position. The judge returns yes, no, or uncertain. It answers all 32 deterministic color, shape, object-absence, and position fixture questions correctly. These elementary checks establish basic pipeline behavior, not accuracy on complex generated scenes.

The primary metric is *confirmed attribute coverage*. For each report and condition, we average the fraction of supported attributes answered yes for the earliest output in that condition; we then average reports equally. Uncertain and unparseable answers are not counted as confirmed, and their counts are retained. Unsupported-literalization probes are reported separately. The questions do not score motion, duration, emotional experience, or whether the original contributor would endorse an image.

For report i, let $c_{i,S}$ and $c_{i,E}$ denote those coverage scores. The primary effect is

$$\widehat{\Delta}_{S-E} = \frac{1}{24}\sum_{i=1}^{24}(c_{i,S} - c_{i,E}). \tag{5}$$

Confidence intervals use 10,000 paired bootstrap resamples of reports; two-sided paired randomization tests enumerate the nonzero report differences exactly (the original Monte Carlo estimate is retained in the result files). The unit of comparison is the report: 24 accounts, not 72 independent people. Intervals remain conditional on this source sample, annotation procedure, image backend, and automated judge.

L Atlas views and comparison examples

An atlas should help someone find a relevant account, understand what is similar and different, and inspect how an illustration was derived. Our offline prototype contains 3,145 source accounts, text search, vividness and duration filters, and side-by-side comparison of original questionnaire responses. The controlled image collection links 84 retained outputs to 24 accounts, with 72 images in the primary matched analysis. A separate blinded interface presents images and source questions for future independent review; no human ratings have been collected.

The account remains the common reference across these views. Study 1 motivates retaining separate reporting channels. Study 2 motivates checking the similarity operation rather than assuming an embedding's probe accuracy guarantees useful retrieval. Study 3 motivates preserving the source span and qualification behind each visual instruction and scoring question. These are design implications of the audits, not measured improvements in user understanding.

Figure 12 illustrates the first two concerns using two source accounts. Their content similarity does not make their separately reported vividness and duration interchangeable. Appendix F documents the schema, and Appendix E retains the earlier source-linked development illustrations and the uncertainty-only record. The development compiler's abstention policy was not used in the matched image study.

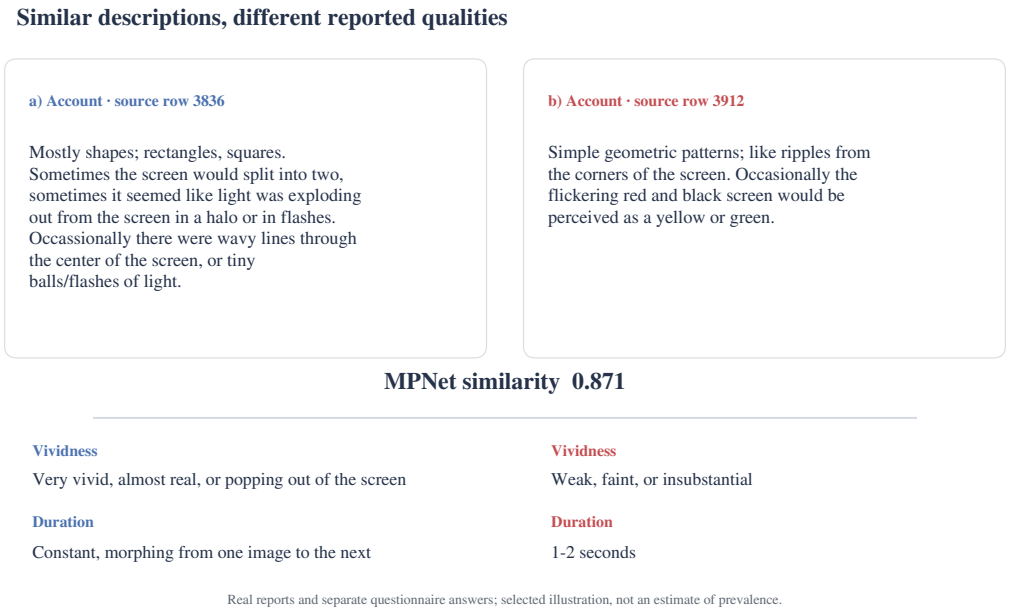


Figure 12: **Similar content, different reported qualities.** Designed example, not an interface screenshot: (a, b) source accounts with their separate questionnaire responses. The pair was selected post hoc for high MPNet similarity among eligible training accounts with matching content category but different vividness and persistence labels. It illustrates a distinction an atlas should expose; it is not an estimate of how often retrieval makes this error.

Table 8: **Three connected questions, three kinds of evidence.** The studies evaluate different parts of the same communication workflow. No single result validates the full path from an experience to another person's understanding.

Question	Evidence	What the result informs
What is reported?	3,145 public accounts; separate questionnaire responses	Which qualities a visual narrative makes predictable.
What meaning is preserved?	672 controlled synthetic contrasts; exploratory probes	Whether wording and readout change access to a distinction.
What does an illustration convey?	24 matched three-condition groups; automated image questions	Visible-detail coverage, prompt-baseline comparison, and limits of the criterion.

Study overview. Table 8 summarizes the different roles of the three studies.

M Post-review exploratory analyses

These analyses were designed after inspecting the original results and an external AI review. They test specific alternative explanations and preserve the original frozen outputs. The protocol was written before the new numerical runs; the original-image reproducibility control described below was added after the achromatic results were available. All analysis code is in `atlas/revision/`; complete outputs are in `results/revision_2026_09_24/`. No new participant data, source narratives, image-generation calls, or representation training were added.

M.1 Development-only threshold and calibration sensitivity

We reuse the original training, development and test groups and feature preprocessing. For each of TF–IDF, MiniLM, MPNet and CLIP text, we refit logistic heads at $C \in \{0.01, 0.1, 1, 10\}$ and select

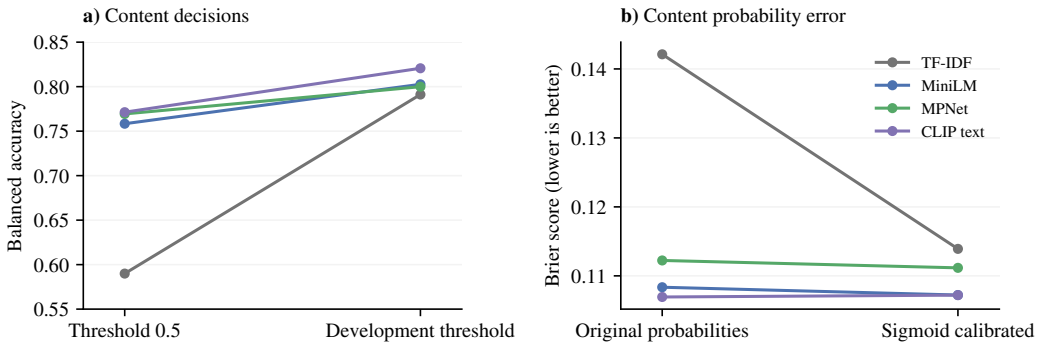


Figure 13: **The content prediction gap depends on the decision protocol.** (a) Test balanced accuracy at the original threshold and a development-selected threshold. (b) Brier error before and after development-fitted sigmoid calibration. Lines connect evaluations of the same fitted encoder/head; they do not indicate longitudinal data or a new representation. These are point estimates from the original held-out sample.

Target	Encoder	BA, 0.5	BA, tuned	Brier	Calibrated
Content	TF-IDF	0.590	0.791	0.142	0.114
	MiniLM	0.758	0.803	0.108	0.107
	MPNet	0.769	0.800	0.112	0.111
	CLIP text	0.771	0.821	0.107	0.107
Vividness	TF-IDF	0.500	0.624	0.235	0.236
	MiniLM	0.573	0.587	0.229	0.221
	MPNet	0.611	0.641	0.225	0.218
	CLIP text	0.580	0.620	0.231	0.222
Persistence	TF-IDF	0.500	0.611	0.231	0.225
	MiniLM	0.605	0.554	0.236	0.225
	MPNet	0.601	0.583	0.233	0.222
	CLIP text	0.602	0.558	0.240	0.226
Affect	TF-IDF	0.500	0.516	0.235	0.234
	MiniLM	0.515	0.510	0.254	0.235
	MPNet	0.523	0.507	0.257	0.235
	CLIP text	0.515	0.517	0.256	0.236

Table 9: **Complete threshold and calibration sensitivity.** BA is balanced accuracy. Each adjustment uses development data only. Some classification scores decrease, and TF–IDF vividness Brier error increases slightly. These post hoc point estimates do not establish that adjustment universally improves a pipeline.

development AUROC, breaking ties by the original loop order. All 16 refitted test-probability vectors exactly reproduce the saved originals in the executed environment.

For classification, we enumerate the finite development ROC thresholds and maximize development balanced accuracy; ties select the threshold closest to 0.5. This changes the decision rule, not the fitted ranking. Separately, we fit a logistic calibration model ($C = 1$, intercept included) to development decision scores and apply it to test scores. It does not use the threshold-selected decisions. Development data are reused for model selection and calibration, making this a pipeline sensitivity rather than an independent validation of a calibration method. The test set determines neither adjustment.

Figure 13 and Table 9 summarize the full sensitivity. These results do not isolate regularization or feature scaling as the sole cause of the original gap. They demonstrate that a comparison using fixed threshold 0.5 can change substantially without changing representations or AUROC. In particular, the majority-class TF–IDF decisions for vividness and persistence become balanced accuracies of 0.624 and 0.611 after threshold selection, whereas affect remains close to chance (0.516).

M.2 SYMMETRIC WORDING AND POLARITY ANALYSIS

Write $t_{c,o,z,w}$ for the existing sentence with content c, operator o, polarity $z \in \{0,1\}$ and wording $w \in \{0,1\}$ (anchor or paraphrase). For each query $q = t_{c,o,z,w}$, its faithful alternative is $p = t_{c,o,z,1-w}$. The query-matched distractor is $n = t_{c,o,1-z,w}$; the matched-alternatives distractor is $n = t_{c,o,1-z,1-w}$. Thus every comparison uses a reworded faithful option, both polarities and both query wordings, while manipulating the distractor's wording. There are $24 \times 7 \times 2 \times 2 \times 2 = 1,344$ comparisons per encoder. Half reuse the original query orientation and half reverse it. Each displayed cell aggregates 96 comparisons, not 96 independent observations.

Cached embeddings and the original text ordering are reused and normalized explicitly. TF–IDF is fitted to the same diagnostic strings without labels, as before. A margin within 10^{-12} of zero is a numerical tie with half credit. The full detail file records every margin and the summary records both polarity and wording breakdowns, with 3,000 content-cluster bootstrap resamples. These intervals condition on the seven template families and are not estimates for unrestricted language. The new comparison fixes query-role asymmetry, not semantic validity or lexical balance across all operators. For example, "on the screen" and "external visual field" differ in specificity.

The complete four-encoder counterbalanced matrix, including TF–IDF, is retained in Figure 14; Figure 4 in the main text brings the neural comparisons and supervised probes together.

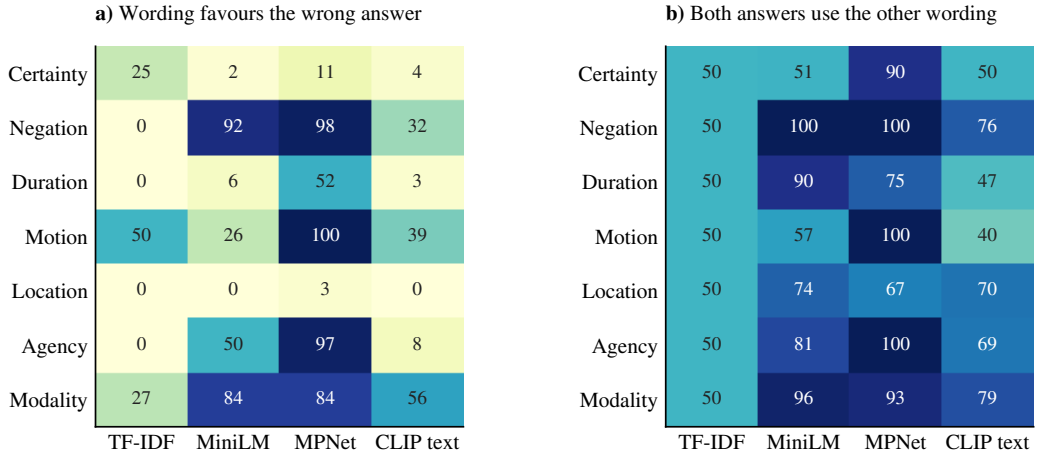


Figure 14: **Full counterbalanced comparison, including the lexical reference.** (a) Query-matched opposite-meaning distractors and (b) matched-alternative wording. Each cell averages 96 comparisons over the same contents, polarities and query wordings. This is the full matrix underlying the neural panels in the main text.

The original paraphrased-distractor scores further illustrate polarity preferences: MiniLM motion is 100%/0%, MPNet duration 0%/100%, and CLIP certainty 83.3%/16.7% across the two polarities. The original plot (Figure 15) and supervised probes (Figure 16) are retained below so that the new diagnostic does not replace their documented design.

M.3 EXHAUSTIVE SOURCE-TO-QUESTION INVENTORY

The inventory includes all 74 annotations and all 69 frozen questions, with report IDs, source spans, complete source text, status tags, and question type. All 74 annotation spans match the source literally. The 65 *reported* annotations and nine other annotations reflect the original authoring, not newly validated semantic classes. Five reported-tag spans contain one of the mechanically screened cues (*like*, *maybe*, *not sure*, *memory/memories*, *probably*, *seemed*, *as if*). Their presence flags an inspection target rather than proving a wrong interpretation.

Table 10 separates the recorded statuses from the question scope. Questions are deliberately restricted to visible content. None asks whether a reader understands the content as remembered, uncertain, or analogical. The memory example demonstrates the distinction: the source contains a sunny childhood beach, the record retains the memory status, and the question asks whether a beach is visible. All

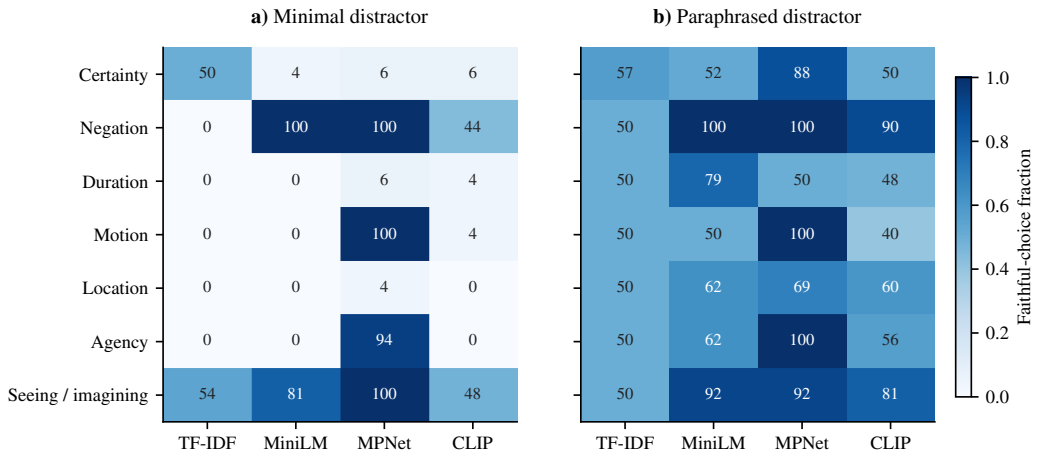


Figure 15: **Original one-direction diagnostic.** (a) Minimal and (b) paraphrased opposite-meaning distractors. Cells average 24 contents and two polarities. Unlike the new Figure 4, queries here always use anchor wording.

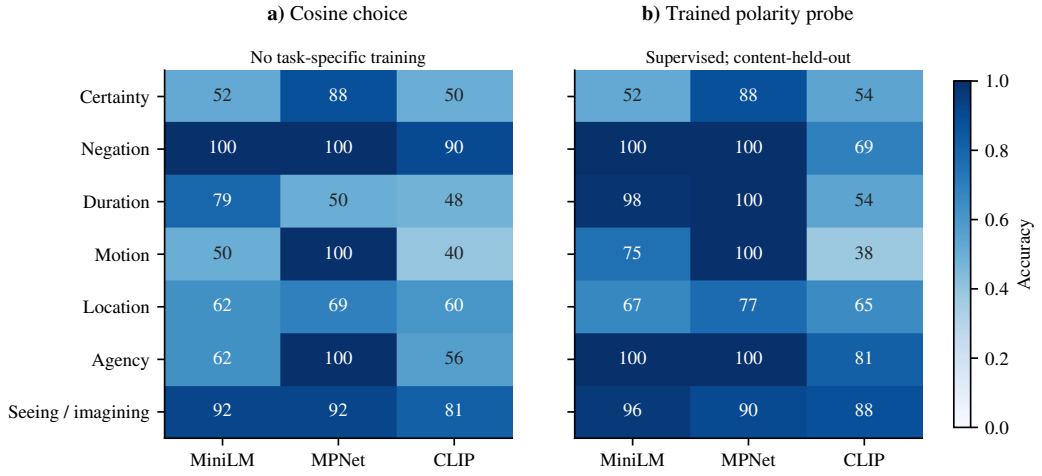


Figure 16: **Original exploratory readout audit.** (a) Cosine faithful-choice accuracy with paraphrased distractors. (b) Content-held-out polarity classification accuracy from trained probes. These are different tasks and supervision regimes; the difference is not a matched method gain.

three images can satisfy that question without establishing the intended interpretation. A caption could supply the status, which is why reader evaluation should assess the complete presentation.

The forest case remains ambiguous: literal trees may be an acceptable way to communicate a forest analogy. We do not label it a hallucination or treat an AI reinterpretation as ground truth. Similarly, a still image's unspecified details are not automatically harmful additions. The intended target is a material unsupported inference, defined with contributors and readers rather than imposed solely by a visual checklist.

M.4 Colour-removal intervention and reproducibility controls

We select every frozen supported question containing a case-insensitive whole-word match for red, green, blue, yellow, pink, orange, or purple. This yields 15 questions in 13 reports. Using all three primary arms gives 45 image–question pairs from 39 unique images (some have two colour questions). No original score or image quality enters selection. The full item manifest records source/output hashes and the exact questions.

Frozen status	Annotations	What the image questions establish
Reported	65	Selected visible content, not full source meaning
Uncertain	4	No direct uncertainty-status question
Analogy	3	Selected object exclusions; no direct analogy-status question
Interpretation	1	Blood-vessel exclusion; no explanatory-status question
Memory	1	Beach content; no memory-status question

Table 10: **Recording a qualification differs from testing its communication.** The nine explicitly qualified annotations occur in eight reports. The status-question classification is an exhaustive AI-authored inspection of the question wording, not human validation of source semantics.

Pillow converts each RGB image to luminance (`L`) and back to RGB, saving lossless PNG at the original resolution. Array assertions confirm equal R, G, and B channels at every pixel. The chromatic predicates in the selected questions are therefore absent under the operational pixel-colour criterion. The intervention preserves geometry but changes luminance relative to the original channels; it is not a perceptual model of colour deficiency or a clinical simulation.

The same local Qwen2.5-VL-3B-Instruct revision and deterministic decoding protocol answer each unchanged question without the source, arm label, or expected answer. All 45 responses are NO; none is uncertain or unparseable. Original saved responses are all YES. After seeing the achromatic results, we add a separate sequential batch rerunning the 45 unmodified originals to check execution reproducibility; its full outputs and timing are retained. All 45 rerun original answers are YES, reproducing the saved answers. This control was not part of the initial exploratory protocol.

Zero observed false yes answers do not establish zero error in the population: the observations share reports, and a bootstrap interval over all-zero outcomes is degenerate. We therefore report the counts and sampling units rather than implying certainty from a collapsed interval. This intervention tests one visible property and one judge. It does not resolve the dark-region judgment, validate the source annotations, or test how an image communicates memory, emotion, duration, or analogy.

M.5 Exact inference and single-answer influence

The original report-level differences contain only one nonzero entry for structured versus prose and two for structured versus direct. Enumerating their 2^1 and 2^2 sign assignments gives two-sided $p = 1.0$ and $p = 0.5$. The original 100,000-draw Monte Carlo value 0.497 for the latter remains in the frozen result file; the manuscript now reports its exact counterpart. Conditional on the observed nonzero differences, even maximally aligned signs cannot yield conventional significance. This is a statement about the realized test, not a prospective power analysis for every possible dataset.

Changing one binary answer changes an arm's macro coverage by $100/(24m_i)$ percentage points, where m_i is that report's number of coverage questions. As an influence scenario, changing the dark-region direct answer to yes reduces structured-minus-direct from 5.56 to 1.39 points. Also changing the forest-tree direct answer to yes reduces it to zero. These hypothetical answer changes are not adjudicated corrections; the original scores and all-output sensitivity remain intact. They show the scale at which criterion or judge disagreements affect this small pilot.

M.6 RETAINED GENERATION SUMMARY AND VISUAL INTERVENTION

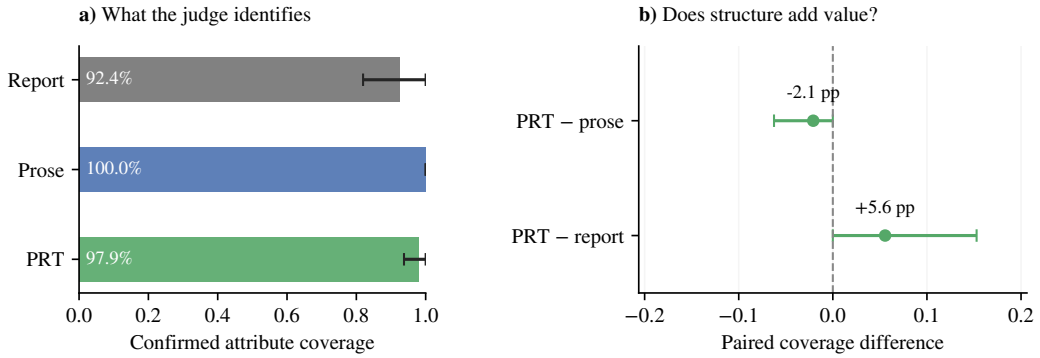


Figure 17: **Visible-attribute coverage under a fixed generation budget.** (a) Mean confirmed coverage from the separate automated judge, weighting reports equally. (b) Paired differences for structured PRT versus evidence-matched prose (primary) and versus the original report (secondary). Intervals are 95% report-bootstrap intervals. Each report contributes one image per condition. These are model judgments of source-linked visible details, not human ratings of experiential fidelity.

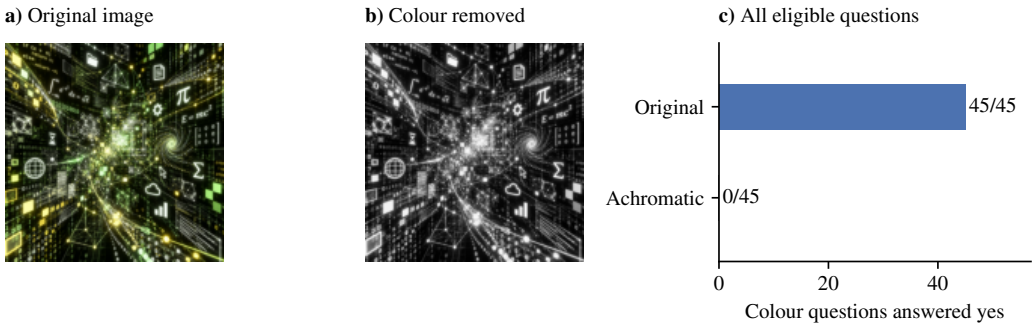


Figure 18: **The judge responds to removed colour, while claim status remains untested.** (a) An original direct-prompt image and (b) its achromatic counterpart; the displayed account is selected by the smallest eligible source ID. (c) Across all 45 eligible image–question pairs, the judge answers yes for the originals and no after colour removal. The 45 pairs share 13 source reports and are not independent participants. This test isolates chromatic evidence, not overall experiential fidelity.

N REAL-REPORT NEAREST-NEIGHBOUR RETRIEVAL

This exploratory analysis was specified after the prediction and synthetic-contrast results, before executing the retrieval run. It evaluates a candidate semantic-search operation for the atlas; the released interface's existing text search is not presented as a validated semantic recommender.

For each of the four targets, the library contains all training records with a valid target and the queries contain all test records with a valid target. Candidate counts are 1,650/1,479/1,406/1,897 and query counts 537/474/450/619 for content/vividness/persistence/affect. We use the cached, normalized MiniLM, MPNet and CLIP text vectors, and TF–IDF with the original training-fitted vocabulary settings. Each query selects the single highest-cosine library vector, with ties resolved by ascending source-record index. The query's target does not enter retrieval; the neighbour's observed binary response is transferred only after selection. No k or distance threshold is selected. Exact-text groups cannot cross the train/test boundary; unknown contributor overlap remains unresolved.

Balanced accuracy gives equal weight to positive and negative query classes. Independently sampling a candidate from the library has expected balanced accuracy 0.5, irrespective of library prevalence. We compute 2,000 bootstrap intervals over query text-hash groups, conditional on the fixed library and representations. All non-content intervals contain 0.5. This does not prove chance-equivalent performance, but it limits the evidence for transferring those qualities through this operation. The four tasks use different valid-target pools and can therefore select different neighbours.

Target	Encoder	Queries	BA	95% interval	Agreement
Content	TF-IDF	537	0.624	[0.583, 0.670]	0.739
	MiniLM	537	0.709	[0.666, 0.753]	0.777
	MPNet	537	0.667	[0.624, 0.713]	0.764
	CLIP	537	0.661	[0.617, 0.705]	0.739
Vividness	TF-IDF	474	0.536	[0.493, 0.583]	0.576
	MiniLM	474	0.539	[0.492, 0.585]	0.565
	MPNet	474	0.526	[0.482, 0.573]	0.549
	CLIP	474	0.532	[0.489, 0.579]	0.555
Persistence	TF-IDF	450	0.535	[0.486, 0.583]	0.553
	MiniLM	450	0.540	[0.494, 0.587]	0.571
	MPNet	450	0.532	[0.483, 0.578]	0.567
	CLIP	450	0.513	[0.465, 0.560]	0.542
Affect	TF-IDF	619	0.497	[0.456, 0.538]	0.522
	MiniLM	619	0.513	[0.474, 0.555]	0.523
	MPNet	619	0.485	[0.446, 0.525]	0.502
	CLIP	619	0.501	[0.461, 0.543]	0.519

Table 11: **Complete held-out nearest-report transfer results.** BA is balanced accuracy of the neighbour's binary questionnaire response relative to the query response; agreement is unweighted accuracy. Intervals resample query text-hash groups and condition on the candidate library. These measure agreement in reported attributes, not human-judged semantic relevance or clinical validity.

Every query index, selected neighbour, cosine and target is retained in `retrieval_details.json`. A descriptive post-retrieval subgroup also records agreement when both reports share the observed coarse content category. This uses the selected pairs without changing their ranking; the coarse category distinguishes complex objects/environments from other valid selections and does not establish identical depicted objects. These subgroup statistics are not a matched causal comparison or the primary retrieval result. The protocol is in `atlas/revision/RETRIEVAL_PROTOCOL.md`; no new images or model-training calls were used.

N.1 A PAIRED REAL-DATA READOUT COMPARISON

After examining the nearest-report results, we compare every model/target combination with its existing development-threshold logistic predictions. Both readouts use the same held-out queries and questionnaire targets; the logistic readout receives supervised fitting while the neighbour rule is untuned. Paired intervals use 2,000 query text-group resamples and are pointwise exploratory intervals without multiplicity correction. This evaluates operation-dependent access to the same frozen representation, not an improved retrieval algorithm.

Target	Encoder	Linear BA	Neighbour BA	Difference [95% interval]
Content	tfidf	0.791	0.624	+0.167 [+0.118, +0.215]
Content	minilm	0.803	0.709	+0.094 [+0.049, +0.140]
Content	mpnet	0.800	0.667	+0.133 [+0.084, +0.185]
Content	clip	0.821	0.661	+0.159 [+0.111, +0.209]
Vividness	tfidf	0.624	0.536	+0.088 [+0.034, +0.146]
Vividness	minilm	0.587	0.539	+0.048 [-0.014, +0.109]
Vividness	mpnet	0.641	0.526	+0.115 [+0.056, +0.172]
Vividness	clip	0.620	0.532	+0.087 [+0.028, +0.146]
Persistence	tfidf	0.611	0.535	+0.076 [+0.018, +0.134]
Persistence	minilm	0.554	0.540	+0.014 [-0.042, +0.074]
Persistence	mpnet	0.583	0.532	+0.052 [-0.008, +0.109]
Persistence	clip	0.558	0.513	+0.045 [-0.021, +0.105]
Affect	tfidf	0.516	0.497	+0.020 [-0.027, +0.069]
Affect	minilm	0.510	0.513	-0.003 [-0.049, +0.041]
Affect	mpnet	0.507	0.485	+0.022 [-0.030, +0.076]
Affect	clip	0.517	0.501	+0.016 [-0.032, +0.066]

Table 12: **Supervised prediction and nearest-report transfer on identical queries.** BA is balanced accuracy. Differences are linear-head BA minus neighbour BA. The linear threshold was selected on development data before this comparison; no test labels tune either readout. Full results prevent a single favourable task from standing in for a uniform effect.

For example, MPNet vividness rises from 0.526 for neighbour transfer to 0.641 for the linear head. Table 12 also shows near-zero or adverse point differences on some tasks. A successful trained readout therefore supports recoverability under that supervision, not reliable transfer of every reported quality through similarity search.