

How to Build a Benchmark That Actually Measures Taste

Xiaofan Wu*, Xi Zeng, Miaoxia Chen,
Peishan Chen, Shuyan Li, Chen Wang, Jiyun Yao, Hanyong Zhong, Jiahao Zhu
Chance AI

Abstract

Expert-aligned “taste” benchmarks should reward craft judgment, not the ability to game the test. We introduce **TasteBench**, a pairwise good-vs-bad benchmark across **visual**, **brand**, **book**, and **tech** items under a frozen rubric with expert-only gold (panel $n \geq 3$; **percent agreement** ≥ 0.60 primary gate; weak engagement labels train/analysis-only), together with a **control protocol** that separates craft judgment from test-gaming. Applying the protocol to a Round-1 pilot (**primary_human_expert**; three biological human experts; $n = 250$ pairs) surfaces and quantifies **three artifacts** that undermine the benchmark’s validity: **(1) position**—the preferred item is slot A with probability 0.956, so an **always-item_a** policy scores 0.956 (fair chance = 0.5); **(2) lexical leakage**—a filename token cue scores 0.912; and **(3) a difficulty ceiling**—craft domains show near-ceiling human agreement (book/brand/tech 1.0), and even under A/B randomization (**eval_assignment_seed7**; **gold_side_A** ≈ 0.47) all five VLM cards reach 1.0 on craft, with only **visual** separating (≈ 0.84 – 0.91 ; a post-hoc hard subset, $n = 107$, leaves craft at 1.0). The protocol—expert-only gold, a weak-label firewall, side randomization *at construction*, filename scrubbing, and hard filters—makes each artifact measurable and, for the two construction artifacts, correctable; we release a reference eval harness and a side-randomization tool, and show the tool removes the position skew ($0.916 \rightarrow 0.495/0.505$). The result is a calibrated view of what the benchmark does and does not yet measure: position and leakage are fixable, whereas the craft-domain ceiling is a pair-difficulty limit that Round-2 must address. We do not evaluate engagement, era/circle conditioning, or calibration here, and we claim no general taste SOTA.

Keywords: vision–language models; evaluation; aesthetic judgment; preference learning; cultural priors; benchmarks

1 Introduction

Large multimodal models are increasingly asked to *judge*: which poster is better designed, which brand system is more coherent, which book cover fits its brief, which API sketch communicates tradeoffs more clearly. These requests sound like taste, but contemporary training mixtures and reward models often optimize for clicks, dwell, shares, and preference data that themselves correlate with engagement. Collapsing “what people engage with” into “what is good” risks mistaking outrage, meme velocity, and recommender amplification for craft—the rival **H3** (engagement-as-taste) in our companion measurement paper.

We take a narrower operational stance. Following the *dubao-taste* program, **taste** is the ability to discriminate good from bad under an explicit rubric of craft, coherence, and appropriateness to brief. TasteBench asks whether VLMs (and text-only ablations) exhibit that discrimination across domains and under social-style conditioning.

*Corresponding author: xiaofan@chance.vision

What this paper is. We build an expert-gold pairwise craft benchmark and a **controlled protocol**, quantify three artifacts that undermine its validity, and show which are correctable. The contribution is the *diagnosis and the controls*, not a leaderboard. We scope the release deliberately: era/circle conditioning and calibration are specified but not exercised, and the engagement-as-taste hypothesis is not tested here. Section 11 gives the exact scope of claims.

Contributions.

1. **A control protocol for trustworthy evaluation.** A small set of controls—expert-only gold, a weak-label firewall, A/B side randomization *at construction*, filename scrubbing, and hard filters on $n_{\text{raters}} \geq 3$ and **percent agreement** ≥ 0.60 —each paired with a measurable audit (Section 7). The protocol is what lets a taste benchmark’s score be read as judgment rather than test-gaming.
2. **Three quantified artifacts it exposes.** A measured *position bias* ($P(\text{pref. in A}) = 0.956$; `always_item_a` scores 0.956), a measured *lexical leak* (`leak_prefer_good_token` scores 0.912), and a measured *difficulty ceiling* (craft saturates at 1.0 for humans and all five VLM cards; only visual separates). The first two are construction artifacts and are correctable; the third is a pair-selection limit.
3. **Released contract and tooling.** Schema tables (with `data_origin`, `pair_group_id`, `cue_type`), `rubric_v1`, versioned prompts (`prompt_vX`), a reference eval harness (`preds` CSV $\rightarrow$ metrics JSON), and a side-randomization tool that demonstrably balances the measured subset ($0.916 \rightarrow 0.495/0.505$).
4. **A decision-relevant finding.** Under the protocol, Round-1 craft pairs do not yet separate models; we isolate *why* (position, leakage, difficulty) and specify the harder, side-randomized Round-2 required—so the negative result is diagnostic, not terminal.

Non-goals. We do **not** test the engagement-as-taste hypothesis in this paper: no engagement labels were scored, so it remains motivation, not a result. We do not claim that TasteBench measures universal beauty, that expert panels are culture-free, or that high TasteBench scores imply product lift in retrieval or recommendation. A planned Know $\rightarrow$ Do follow-on is locked as **offline Do-slot protocol eval** (plan Option A in the companion track): TasteBench-aligned VLM scores as Do policy priors vs random / position / engagement-style foils, measured by agreement with the *same* expert gold used here; hard-subset stress as a first cut; **no InternVL3 weights claim**; live A/B / causal CTR out of scope. **No product results** are claimed here.

2 Related Work

Preference learning and pairwise judgment. Bradley–Terry and related pairwise models [Bradley and Terry, 1952] underpin modern preference optimization and human-feedback pipelines [Christiano et al., 2017, Ouyang et al., 2022, Rafailov et al., 2023]. TasteBench adopts pairwise preference as the primary task because it matches how craft experts often adjudicate close calls; primary pair gold is majority vote among expert raters, with Bradley–Terry reserved for global ranking / analysis / sensitivity when the within-domain pairwise graph is dense enough (mean degree ≥ 3).

Aesthetic assessment and “quality” datasets. Computer vision has long predicted aesthetic scores and popularity from photography and design corpora [Murray et al., 2012, Kong et al., 2016]. Many such datasets operationalize quality via mean opinion scores that partly track appeal or engagement. TasteBench positions itself against that collapse: engagement may

correlate with expert judgment in some slices and diverge sharply in others (controversy, rage-bait, SEO packing). TasteBench is designed to treat engagement-trained predictors as baselines and negative controls, not oracle labels; this release does not exercise them (Section 11).

Cultural capital, habitus, and circle priors. Sociological accounts of taste emphasize acquired dispositions and field-specific recognition [Bourdieu, 1984]. We do not claim a full operationalization of Bourdieu; we only encode **era** and **circle** as *conditioning instructions* under which experts and models must judge fit to a stated prior—so that “good for Y2K poster design” can differ from “good for 2020s minimal product UI” without treating platform traffic as truth.

VLM evaluation beyond recognition. Recent multimodal benchmarks such as MMMU, MMBench, and SEED-Bench stress perception, OCR, chart reasoning, and instruction following [Yue et al., 2024, Liu et al., 2023, Li et al., 2023]. Fewer isolate *normative craft discrimination* under explicit non-virality rubrics. TasteBench targets that gap with a cross-domain meta-skill framing: the same judgment type (good vs bad under brief) on different surfaces.

Recommender confounding. Observational feeds confound preference with position and recommendation [Joachims et al., 2017]. The companion paper studies that identification problem for *people*. TasteBench instead evaluates *models* against expert gold, while allowing engagement signals only as weak train-time material—so success cannot be achieved by reconstructing the engagement head by construction.

Calibration. Expected calibration error [Guo et al., 2017] is our intended secondary metric; it is specified in the eval contract but *not reported* in this release (Section 11).

3 Benchmark Design

3.1 Goal and task typology

TasteBench evaluates whether a model can discriminate **good vs bad** under `rubric_v1`, across domains, optionally conditioned on era/circle. Tasks:

1. **Pairwise preference (primary).** Given items A and B (and optional condition), choose which is better under the rubric. **Primary pair gold = majority vote** among expert raters. **Bradley–Terry** is used only when the within-domain pairwise graph has **mean degree** ≥ 3 , and only for global ranking / analysis / sensitivity tables—it **does not replace** primary test labels by default.
2. **Absolute quality (secondary).** Score an item on $[0, 1]$ Likert-calibrated bands; used for calibration analyses and the companion paper’s τ family, not as the primary TasteBench leaderboard number.
3. **Conditional taste.** The same content pair under different `condition_era` / `condition_circle` instructions, linked by `pair_group_id` with `cue_type` $\in \{\text{null, correct, misleading}\}$. This measures sensitivity to stated priors—not engagement prediction.

3.2 Domains (v1)

3.3 Schema

Four core tables:

Domain	Example items	Model inputs
visual	posters, UI stills, photography	image (+ optional short caption)
brand	logo/identity pairs, packaging	image + brand-agnostic brief
book	covers + blurbs, or passage craft	image and/or text
tech	API snippets, architecture diagrams, product UX	image and/or text

Table 1: Four Round-1 domains. Cross-domain transfer is intentional: “can tell good from bad” is treated as one meta-skill with different surface forms.

- **bench_items:** `item_id`, `domain`, `uri_or_payload_ref`, optional `era_tag` / `circle_tags`, required `license` (redistribution-OK assets only), required `data_origin`.
- **bench_pairs:** `pair_id`, `domain`, `item_a`, `item_b`, optional `condition_era` / `condition_circle`, optional `pair_group_id` (family for correct vs misleading cue), optional `cue_type` $\in \{\text{null, correct, misleading}\}$, `difficulty` $\in \{\text{easy, medium, hard, ambiguous}\}$, `split` $\in \{\text{train, dev, test}\}$, optional `cached_agreement` / `n_raters`, required `data_origin`.
- **bench_gold:** `pair_id`, `preferred_id` (pair-grain **expert majority**; primary test labels), `agreement` (**percent agreement**, primary), `n_raters`, `rubric_id`, `data_origin`. The companion paper exposes the same aggregates via `label_aggregates` (`label_grain` $\in \{\text{item, pair}\}$, `preferred_id`, `exclude_primary`). Optional BT ranks (within-domain mean degree ≥ 3) are analysis/sensitivity only—not default `preferred_id`.
- **bench_weak (optional):** `pair_id`, `weak_preferred_id`, `weak_source`, `use_as` $\in \{\text{alignment_only, analysis_only}\}$ —**never** test labels; `data_origin`.

Label policy (frozen). (1) Test = expert gold only via `bench_gold` / `label_aggregates`. (2) Weak engagement never enters test metrics (train alignment / analysis only). (3) Simulation stress tests are allowed only with `data_origin=simulation` and methods disclosure—never silently merged into primary claims. (4) Eval **hard-filters** `n_raters` ≥ 3 and percent **agreement** ≥ 0.60 (inclusive); drops `difficulty=ambiguous` / `exclude_primary` even if tags are wrong.

3.4 Era and circle conditioning (specified, not exercised in this release)

Conditions are *instructions*, not hidden features of popularity. Raters and models are told, when a condition is present, to judge fit to that prior; if the condition conflicts with craft, free-text notes are allowed, but the primary label follows rubric + condition. Conditional variants of the same content pair share `pair_group_id`; `cue_type` $\in \{\text{null, correct, misleading}\}$ marks unconditioned, correctly matched, or deliberately wrong era/circle prompts. Misleading-cue ablations would support the condition-sensitivity metric.

Important scope note. Although the schema materializes 10 `null+correct+misleading` families (20 conditioned rows), these rows are `split=dev`, carry **no expert labels**, appear in **zero** rows of the primary gold, the eval assignment, or any model prediction, and were never rendered into a cue banner in `prompt_v1`. **Condition sensitivity is therefore a specified design element that is not evaluated in this release.** We report no condition-sensitivity numbers (Section 11).

Circle tags vs. membership. TasteBench uses item-side `circle_tags` (and `condition_circle`) as *condition cues*. User circle membership (`user_circles`) is the companion paper’s primary H2 construct—optional cross-reference only; it is not required to score TasteBench pairs.

3.5 Relationship to companion work

The companion measurement paper quantifies *human* taste as the shape of high-quality exposure memory (τ / half-life / `user_circles` H2 live there) and pre-registers H3 (engagement-as-taste) as a rival; both efforts share the expert-gold path (`label_aggregates` $\leftrightarrow$ `bench_gold`) under the same gold/weak firewall. TasteBench supplies the model-facing expert-gold measurement. A companion Know $\rightarrow$ Do track is locked as **offline Do-slot protocol eval** (plan Option A); a first $K=2$ artifact exists in that track (same expert gold; VLM predictions as Do scores vs random / position / construction foils; H3 engagement $\neq$ quality as negative-control framing; hard-subset stress). Live product A/B / causal CTR remains **out of scope**—**this paper cites no companion-track cells as results**.

4 Annotation Protocol

4.1 Rubric (`rubric_v1`)

Experts answer: given the brief (or domain default), which option is **better crafted** / **more coherent** / **more appropriate**? Explicitly *not* judged: virality, follower count, controversy heat, or platform of origin alone. Domain cues instantiate the same meta-skill:

- **visual**: composition, hierarchy, restraint, technical finish
- **brand**: distinctiveness, consistency, brief-fit (not logo fame)
- **book**: prose craft / cover-content fit / editorial judgment (task-dependent)
- **tech**: clarity of API/UX/architecture tradeoff communication; correctness over hype

Absolute anchors (0–1): clear fail (0.0–0.3), competent/generic (0.4–0.6), strong (0.7–0.85), exemplary (0.85–1.0). Pairwise instructions mirror these bands. Guideline edits bump the version id; this release freezes `rubric_v1`.

4.2 Panel process and inclusion rules

Round-1 primary panel. Raters `human_expert_01` / `human_expert_02` / `human_expert_03` are **biological human experts**; any agent protocol used for packet delivery is a **communication channel only** (user-confirmed). Promoted gold path: `bench_gold_primary_human.csv` (`gold_tier=primary_human_expert`; all rows `exclude_primary=false`). The superseded archival file `bench_gold_provisional_agent_panel.csv` is **not** the current analysis tier.

1. At least **three** independent raters per item/pair; **no engagement statistics** shown.
2. Aggregate: **majority vote = primary pair gold** (`preferred_id`); mean for absolute scores. **Bradley–Terry** only if within-domain pairwise graph **mean degree** ≥ 3 , and only for global ranking / analysis / sensitivity tables—**does not replace** primary test labels by default (sparse graphs stay majority-only).
3. **Primary agreement gate: percent agreement ≥ 0.60 (inclusive)** on the discrete label (pairwise `preferred_id` or binned absolute). Below gate *or* $n < 3 \rightarrow \text{difficulty=ambiguous}$ and `exclude_primary=true`, **hard-filtered from the primary test metric** (still retained for analysis of hard cases). **Secondary (appendix only):** Krippendorff’s α (nominal) [Krippendorff, 2018]—does not replace the primary gate. Eval must **hard-filter** on $n_{\text{raters}} \geq 3$ and percent agreement ≥ 0.60 , not trust difficulty tags alone. (With $n = 3$, percent agreement takes only the values $\{1, 2/3, 1/3\}$, so the ≥ 0.60 gate is equivalent to a $\geq 2/3$ majority.)
4. Companion defaults ($\tau = 0.70$ sweep $\{0.60, 0.70, 0.80\}$; recency half-life **90 days**) apply to human exposure / **memory_shape in the companion paper**—**cross-ref only**; they are not TasteBench’s primary pairwise metric and are not re-frozen here.

4.3 Weak supervision firewall

High-interaction platform subsets may yield `bench_weak` rows for contrastive warmup or disagreement analysis (`use_as=alignment_only` / `analysis_only`). They are barred from test labels and from defining leaderboard success. **Test gold = expert only** via `bench_gold` / `label_aggregates`. Engagement-as-taste remains H3: report correlation and divergence (controversy, amplification, outrage), never silently substitute. *In the Round-1 pilot no weak/engagement pairs were evaluated; the firewall is a contract, not an exercised comparison.*

4.4 Stub corpus (non-empirical)

The repository ships placeholder CSVs to exercise the contract: 50 items across four domains, 50 pairs (30 test / 10 dev / 10 train), 19 gold pairs meeting $n \geq 3$ and percent agreement ≥ 0.60 , and 10 weak train-only rows. All stub tables carry `data_origin` (typically `simulation` for wiring). Assets are CC-BY-4.0-placeholder paths—**not a licensed release corpus**. **No scientific claims are made from stubs; stub/toy numbers are never presented as real results.**

5 Metrics and Baselines

5.1 Metrics

Let $\mathcal{T}$ be the primary test set: gold pairs with `split=test`, hard-filtered to $n_raters \geq 3$ and **percent agreement** ≥ 0.60 (inclusive), dropping `difficulty=ambiguous` / `exclude_primary` (even if tags disagree).

- **Pairwise accuracy (primary).** Fraction of pairs where `pred_preferred_id` equals `preferred_id`. Reported overall and by domain.
- **ECE.** If the model emits confidence in $[0, 1]$, compute expected calibration error over correctness indicators (default 10 bins) [Guo et al., 2017]. *Not reported in this release (no confidence emitted).*
- **Condition sensitivity.** Compare accuracy by `cue_type` within each `pair_group_id`—correct vs misleading vs null. *Not reported in this release (no conditioned rows scored).*
- **Cross-domain spread (zero-shot).** Because the Round-1 cards are used zero-shot (no per-domain adaptation), we report the *spread* of per-domain accuracy ($\max - \min$) as a consistency diagnostic rather than a transfer matrix (Table 10). A true transfer gap would require domain-adapted runs.
- **Explicit non-metric.** Engagement / click prediction accuracy is **not** a TasteBench success criterion.

5.2 Baselines

1. **Strong VLMs** with versioned `prompt_vX` templates (pairwise forced choice + optional confidence).
2. **Text-ablated** counterparts (captions/briefs only) to separate visual from linguistic priors.
3. **Retrieval / aesthetic-score proxies** from prior aesthetic literature [Murray et al., 2012, Kong et al., 2016].

4. **Engagement-trained models as negative controls**—systems optimized for likes, saves, or watch time. Competitive engagement prediction with *poor* expert-aligned TasteBench accuracy supports the dissociation thesis; the reverse pattern would favor H3 in that slice and must be reported honestly.

Random / majority-class chance levels are reported per split imbalance. Oracle predictions (copying gold) validate the eval harness only. Empirical non-VLM floors on primary gold ($n = 250$) are reported in Section 8.2; the high `always_item_a` accuracy is a construction artifact, not a ceiling.

5.3 Eval contract

- **Input:** predictions CSV with columns `pair_id`, `pred_preferred_id`, optional `confidence`.
- **Gold path:** merge to `bench_gold` only (never `bench_weak`); filter `split=test`; **hard-filter** $n_raters \geq 3$ and percent `agreement` ≥ 0.60 (inclusive); drop `difficulty=ambiguous` / `exclude_primary`.
- **Output:** metrics JSON including `n_eval`, `pairwise_acc`, `by_domain`, optional `ece`, accuracies by `cue_type` / `pair_group_id` when present, and a policy string restating the weak-label firewall and `data_origin` disclosure.

A reference implementation of this contract (reconstructed from the frozen spec; *not* yet reconciled with the original Round-1 runner) ships at `scripts/tastebench_eval.py`; it computes pairwise accuracy, by-domain accuracy, ECE (when a `confidence` column is present), and by-`cue_type` slices. A confidence-eliciting template `prompt_v2` (and a conditioned variant `prompt_cue_v1`) is provided so ECE and cue sensitivity become computable on the next run, even though they are not reported here.

6 Protocol

Prompting. All submissions declare `prompt_vX`, model id/revision, decoding settings, and whether images, briefs, or both were provided. Condition strings, when used, must match schema enums or be declared as adversarial/misleading ablations.

Splits. Train may include weak labels for alignment; dev is for selection; test metrics use expert gold only. Rows with `data_origin=simulation`, if any, are disclosed in methods and reported separately—never silently merged into primary claims.

Evaluation / ranking rules (intended). Primary sort key: pairwise accuracy on $\mathcal{T}$. Secondary: ECE (lower better), mean condition-sensitivity score, and cross-domain spread (smaller is more consistent, with per-domain specialists reported separately). Engagement metrics may appear in an appendix as diagnostics, never as the ranking key.

Reproducibility and release status. The intended release includes the schema, rubric one-pager, rater guidelines, anonymized gold aggregates, the eval script, the versioned prompts, and the raw prediction files. **This preprint ships the aggregate tables, per-card scores, two derived diagnostics (side-balance, cross-domain spread), and five figures; it also ships a reference eval harness and the `prompt_v2` template.** The raw `preds_*.csv` from the Round-1 run and the eval assignment are enumerated in `RELEASE_MANIFEST.md` and must accompany the camera-ready for end-to-end reproduction. We flag this rather than overstate reproducibility.

#	Control	Neutralizes	Audit
C1	Expert-only gold; weak labels train/analysis-only	virality-as-taste	gold provenance (data_origin)
C2	Hard filter: $n_raters \geq 3$ AND percent agreement ≥ 0.60	noisy / ambiguous labels	agreement distribution
C3	Randomize A/B side at construction	position bias	side-balance audit (Table 7)
C4	Scrub predictive tokens from filenames / IDs	lexical leakage	leak baseline (Table 6)
C5	Report difficulty; reject or replace saturated domains	difficulty ceiling	per-domain spread (Table 10)

Table 2: The control protocol. Each control is paired with an audit that makes the artifact visible before the score is trusted. C1–C2 are enforced by the released eval harness; C3 is implemented by `randomize_pair_sides.py`; C4 is a data-hygiene step; C5 is a reporting requirement.

Toy harness check (not a result). On the stub gold set ($n = 19$), a chance-like prediction file yields pairwise accuracy ≈ 0.42 (toy) with ECE ≈ 0.13 (toy); an oracle file yields accuracy 1.0 (harness sanity). These figures validate plumbing only (Appendix B).

7 A Control Protocol for Expert-Gold Pairwise Evaluation

The protocol is the paper’s primary artifact. It is a small set of controls, each paired with a measurable audit, such that a benchmark’s pairwise score can be interpreted as craft judgment rather than test-gaming. None of the controls is novel in isolation; what makes the score trustworthy is the *combination*, together with the requirement that each artifact be audited and reported alongside accuracy.

How to read a TasteBench score. A score is meaningful only together with its audits: report the side-balance audit (C3), the leak-baseline score (C4), and the per-domain spread (C5) next to pairwise accuracy. In our Round-1 data, C3 and C4 show two trivial policies scoring 0.956 and 0.912 (Section 8.2)—without them, almost any leaderboard number is uninterpretable.

8 Results

8.1 Primary panel and gold (descriptive)

Disclosure (required). Round-1 primary gold path: `labeling/round1/aggregates/bench_gold_primary_human.csv` (tier `primary_human_expert`). Rater IDs `human_expert_01/02/03` are **biological human experts**; the agent protocol is a **communication channel only**. Every primary gold row is tagged `gold_tier=primary_human_expert`, `data_origin=study_human_expert`, and `exclude_primary=false`. The older provisional agent-panel file is **superseded archival only**.

Primary gold is **ready for controlled evaluation** under the expert-only test policy. Near-ceiling agreement may still reflect easy good/bad craft contrasts; it must not be cited as a VLM performance ceiling. Intent $QC \neq \text{gold}$.

Quantity	Value
Label rows	750
Gold pairs (primary)	250
Raters	human_expert_01, 02, 03 ($n_{\text{raters}} = 3$ on all pairs)
Primary-eligible	250 / 250
Ambiguous	0
Unanimous triples	247 / 250
Rubric	rubric_v1
Intent QC match	1.0 on 225 pairs (construction check only)

Table 3: Primary panel / gold coverage. Intent QC is a construction check, not gold and not a model score.

Domain	Pairs	Mean agreement
book	60	1.0
brand	60	1.0
tech	50	1.0
visual	80	≈ 0.9875

Table 4: Domain mean agreement. Craft domains are unanimous; only three visual pairs are not.

8.2 Non-VLM baselines (primary gold; $n = 250$)

These are **non-VLM** floors / diagnostics only—**not** a model leaderboard and **not** VLM numbers.

Required disclosure. The high accuracy of `always_item_a` (0.956 overall; 1.0 on book/brand/tech) is a **dataset construction artifact**: preferred / “good” items are often placed as `item_a` in the Round-1 pair files. It is **not** a model performance ceiling and must not be cited as one. Symmetrically, `always_item_b` (0.044) is the complementary artifact.

Fair floors. Treat `random_seed42` (0.552) and `hash_coin` (0.5) as the fair chance / non-informative floors.

Filename-cue diagnostic only. `leak_prefer_good_token` (0.912) exploits a self-authored filename token cue; it is **not** a fair baseline—report only as a leakage diagnostic.

Recommendation. Future annotation / construction rounds should **randomize A/B side** (preferred as `item_a` vs `item_b`) so side-biased trivial policies cease to look like ceilings. Tooling is provided (`scripts/randomize_pair_sides.py`).

Side-balance audit and the effect of randomization. Table 7 makes the position artifact quantitative. On the full 250-pair gold the preferred item sits in slot A with probability 0.956 (book/brand/tech 1.0; visual 0.8625). The 107-pair hard subset, whose per-pair slots are available, shows the same skew (0.916 overall). Applying the released side-randomization tool to that subset (seed 42) balances the sides to 0.4953/0.5047, turning the position-only floor from a strong-looking baseline into unbiased chance. Re-deriving the full 250 requires the Round-1 pair file; the protocol is to randomize side *at construction* in Round-2.

8.3 Round-1 VLM results (the ceiling in practice)

Protocol. Provider = OpenRouter chat completions; assignment = `eval_assignment_seed7.csv` ($n = 250$, seed = 7; **gold_side_A rate** ≈ 0.47 —A/B randomized so the construction artifact of Section 8.2 is mitigated for VLM scoring); prompt = `prompt_v1`; model map locked at card version v1.2. Pairwise accuracy overall and by domain against `bench_gold_primary_human.csv`.

Agreement	Pairs
1.0	247
0.6667	3

Table 5: Agreement distribution. The three non-unanimous pairs are all Met **visual** (pair_013, _015, _016).

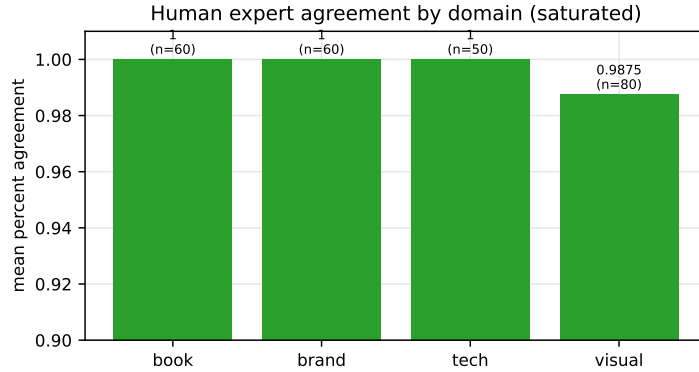


Figure 1: Human expert mean percent agreement by domain. Craft domains are saturated at 1.0; the Round-1 contrasts are easy.

Required disclosure — InternVL3-8B runtime substitute. The locked map intended `nvidia/nemotron-nano-12b-v2-v1:free` for the InternVL3-8B card; the free tier stalled / rate-limited ($\approx 12/250$ in ≈ 3 min), so the finished predictions for that card are from `google/gemma-3-12b-it`. Mark as **runtime substitute**. Do **not** claim InternVL3 weights were evaluated. Four other cards ran on their locked OpenRouter IDs.

Limitations of this Round-1 slice. Book / brand / tech = 1.0 for all five cards likely reflects easy craft contrasts plus near-ceiling human IRR (Section 8.1), not universal mastery. **Visual** is where models separate (≈ 0.84 – 0.91). Compare to fair chance `hash_coin` = 0.5. Do **not** claim general SOTA taste without harder pairs, contested items, or broader domain coverage. ECE and condition sensitivity are **not reported** for this run (no confidence column $\rightarrow$ ECE not computable; cue triples are 0 in gold/preds $\rightarrow$ cue Δ acc not scorable); zero-shot cross-domain spread is reported in Table 10. **All accuracies are point estimates and we do not significance-test model differences:** at $n = 250$ ($n = 50$ – 80 per domain) the 95% binomial half-width is roughly ± 0.03 (± 0.08 per domain), so small inter-card gaps (e.g. gpt-5.6-terra and claude-sonnet-5 differ by a single pair) are within noise.

Post-hoc craft hard subset (stress slice; not new gold). Composition: **visual 40 / book 27 / brand 20 / tech 20** = 107 pairs; every row keeps primary-gold `preferred_id` with `include_in_eval=true`. Primary gold is unchanged. **Honest shortfall:** because Round-1 craft VLMs sit at ceiling on book/brand/tech (accuracy = 1.0), the brand/tech hard slices are **low-expert-confidence only** (weak discrimination; padded to $n = 20$); book uses filename-cue fails + low confidence—**not** VLM-error hardness. Scored post-hoc from existing predictions; because the visual slice was selected using the same models’ errors, its numbers are descriptive rather than an independent estimate of skill:

Zero-shot cross-domain spread. Because all cards are used zero-shot, the honest generalization number is the spread of per-domain accuracy (max – min), not a transfer matrix:

Roster context. Explicitly not in the main table: GPT-4o / 4o-mini; Claude 3.5–4.x as primary; Gemini 1.5 / 2.5 as primary; Qwen2 / Qwen2.5-VL; classic LLaVA-OneVision-7B. Stub/toy harness checks (Appendix B) remain **plumbing validation only** and are **not** leaderboard entries.

baseline	n	acc	book	brand	tech	visual
always_item_a	250	0.956	1.0	1.0	1.0	0.8625
always_item_b	250	0.044	0.0	0.0	0.0	0.1375
random_seed42	250	0.552	0.5833	0.55	0.6	0.5
hash_coin	250	0.500	0.5667	0.45	0.44	0.525
leak_prefer_good_token	250	0.912	0.8333	1.0	1.0	0.85

Table 6: Non-VLM floors on primary gold. `always_item_a` high accuracy is a construction artifact, not a ceiling; fair floors are `random_seed42` / `hash_coin`; `leak_prefer_good_token` is a leakage diagnostic only.

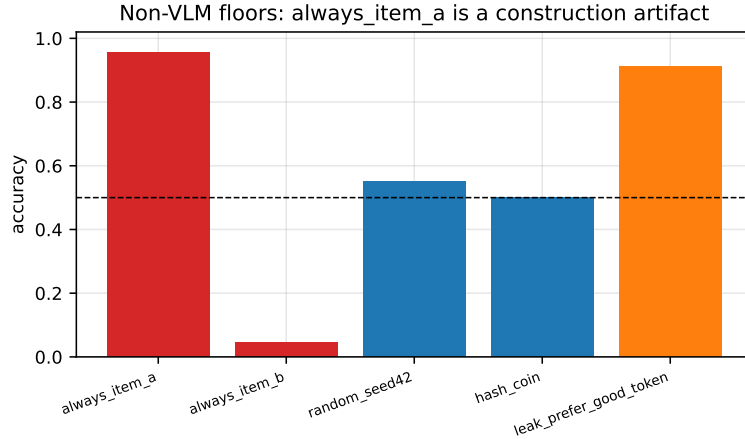


Figure 2: Non-VLM floors on primary gold. `always_item_a` and `leak_prefer_good_token` are artifacts; fair floors are `random_seed42` / `hash_coin`.

9 Beyond TasteBench: recurring artifacts and a checklist

The three artifacts are not specific to our pairs. Any pairwise “good-vs-bad” benchmark assembled from curated assets with semantic identifiers and an expert panel can face them, so the audits generalize:

Why only visual separates. The craft domains saturate because the Round-1 pairs are easy (Section 8.1)—*not* because of position: the VLM eval used a side-balanced assignment (`gold_side_A` ≈ 0.47), under which a position-only policy scores ≈ 0.47 , far below the observed 1.0. Position bias affects the *trivial-baseline* layer of the original construction; the craft saturation is a difficulty limit. Visual is the only domain whose pairs carry a genuine accuracy gradient.

Checklist for building an expert-gold pairwise benchmark.

1. **Randomize A/B side at construction** and report the side-balance audit; a position-only policy must score $\approx$ chance.
2. **Scrub predictive tokens** from filenames, IDs, and paths; a filename-only baseline must score $\approx$ chance.
3. **Freeze expert-only gold** and keep any engagement signal train/analysis-only.
4. **Hard-filter** on $n_{raters} \geq 3$ and percent agreement ≥ 0.60 ; report the agreement distribution.
5. **Report per-domain accuracy**, and treat a saturated domain (all models at ceiling) as *unmeasured*, not as solved.

scope	domain	n	P(pref. is A)	P(pref. is B)
full 250	all	250	0.9560	0.0440
full 250	book/brand/tech	170	1.0	0.0
full 250	visual	80	0.8625	0.1375
hard 107	all	107	0.9159	0.0841
hard 107	visual	40	0.775	0.225
hard 107 (side-randomized, seed 42)	all	107	0.4953	0.5047

Table 7: Side-balance audit. The preferred item sits in slot A far more often than chance; the released randomization tool (`randomize_pair_sides.py`) restores balance on the subset with per-pair slots. Sources: `tastebench_baselines_summary.csv`, `craft_hard_subset.csv`.

model card	runtime openrouter_id	acc	book	brand	tech	visual
gpt-5.6-terra	openai/gpt-5.6-terra	0.960	1.0	1.0	1.0	0.875
claude-sonnet-5	anthropic/claude-sonnet-5	0.956	1.0	1.0	1.0	0.8625
gemini-3.7-flash	google/gemini-3.7-flash	0.972	1.0	1.0	1.0	0.9125
Qwen3-VL-8B	qwen/qwen3-vl-8b-instruct	0.948	1.0	1.0	1.0	0.8375
InternVL3-8B (card)	google/gemma-3-12b-it	0.968	1.0	1.0	1.0	0.900

Table 8: Round-1 VLM results. Craft domains are at ceiling for all cards; only visual separates. InternVL3-8B is a disclosed runtime substitute.

6. **Report the trivial-policy ceiling** (`always_item_a`, filename leak) next to model scores.
7. **Release the audits and the harness**, not only the accuracy table.

Applicability to platform-scale and social-media data. The controls are agnostic to where a corpus comes from, so they transfer directly to platform-scale and social-media data. A content platform is in some ways *better* positioned than this academic pilot: it can assemble large expert-gold sets, and it owns the engagement logs needed to run the negative control we could not (Section 11)—mapping where engagement and expert quality diverge. The discipline is unchanged: engagement may inform analysis but must not enter the gold, sides must be randomized at construction, and identifiers must be scrubbed before release. Applied this way, the protocol turns a private corpus into a trustworthy craft-judgment benchmark; applied with engagement as the answer key, it reproduces exactly the artifacts above.

10 Analysis Plan

Pre-registered analyses (aligned with frozen defaults):

1. **Discrimination vs engagement.** Correlate model preference with expert gold vs with weak engagement labels on the same pairs; report regions of agreement and conflict. (*Not exercised in Round-1.*)
2. **Difficulty slices.** Accuracy on easy/medium/hard gold pairs; separate reporting for ambiguous (excluded) pairs. The post-hoc craft hard subset ($n = 107$) is a disclosed stress slice—**not** new gold.
3. **Prior sensitivity.** Within `pair_group_id`, does `cue_type=correct` improve expert-aligned accuracy vs misleading/null more for VLMs than for engagement-trained controls? (*Not scorable in Round-1: no conditioned rows scored.*)
4. **Modality ablations.** Image-only vs text-only vs both, by domain.

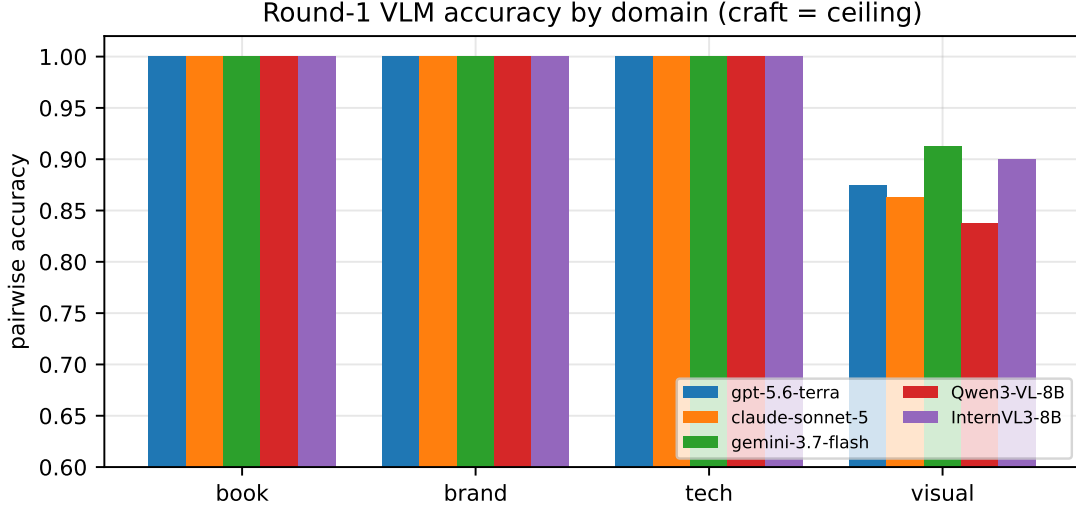


Figure 3: Round-1 VLM accuracy by domain. Craft domains are at ceiling for every card; only visual separates.

model card	hard n	hard acc	visual (40)	book (27)	brand (20)	tech (20)
gemini-3.7-flash	107	0.9346	0.825	1.0	1.0	1.0
InternVL3-8B (card/gemma)	107	0.9252	0.800	1.0	1.0	1.0
gpt-5.6-terra	107	0.9065	0.750	1.0	1.0	1.0
claude-sonnet-5	107	0.8972	0.725	1.0	1.0	1.0
Qwen3-VL-8B	107	0.8785	0.675	1.0	1.0	1.0

Table 9: Post-hoc hard stress subset. Craft domains remain at 1.0, consistent with the ceiling shortfall. Cite as stress disclosure only; do not replace full primary-gold metrics.

5. **Error taxonomy.** Manual codes: brief mismatch, era anachronism, fame/brand recognition bias, toxicity/virality bias, technical incorrectness.
6. **Robustness.** Prompt paraphrase sets under the same `prompt_vX` family; seed variance for sampling decoders.
7. **Human ceiling.** Panel self-consistency and held-out expert agreement as soft upper bounds—not proof of objectivity.

11 Limitations

- **Expert panels are situated.** Rubric-based craft judgment still reflects who is recruited; era/circle tags cannot exhaust cultural space.
- **Agreement threshold.** Excluding percent agreement < 0.60 improves metric stability but may discard theoretically interesting contested taste; α remains appendix-only.
- **Domain coverage.** Four domains are illustrative, not exhaustive (music, fashion film, cuisine, etc., deferred).
- **Conditioning not evaluated.** Era/circle prompts (`cue_type` on `circle_tags` / era cues) are specified but **not exercised** in this release; no condition-sensitivity number is reported. No causal claim about lived subcultural membership is made.

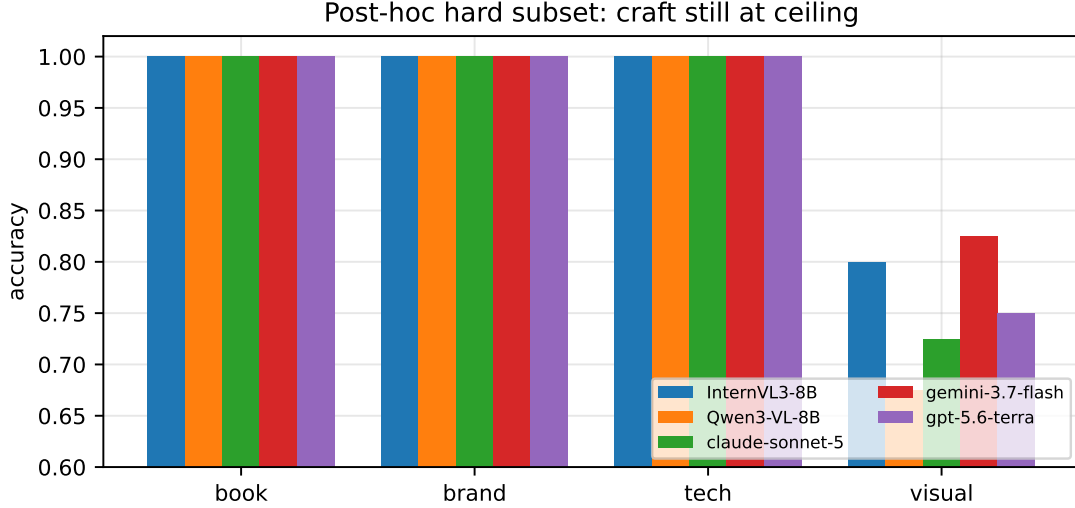


Figure 4: Post-hoc hard subset: craft domains still at 1.0; only visual moves.

model card	book	brand	tech	visual	spread
gemini-3.7-flash	1.0	1.0	1.0	0.9125	0.0875
InternVL3-8B (gemma)	1.0	1.0	1.0	0.900	0.1000
gpt-5.6-terra	1.0	1.0	1.0	0.875	0.1250
claude-sonnet-5	1.0	1.0	1.0	0.8625	0.1375
Qwen3-VL-8B	1.0	1.0	1.0	0.8375	0.1625

Table 10: Zero-shot cross-domain spread (max domain accuracy – min). Sources: `vlm_main-table.csv`, `cross_domain_spread.csv`.

- **Weak labels not exercised.** Platform engagement filters are noisy and platform-specific; they exist only for alignment/analysis, and none were evaluated in Round-1.
- **No product causal claim.** Mapping TasteBench competence $\rightarrow$ Know $\rightarrow$ Do is offline Do-slot protocol eval (plan Option A) in the companion track—same expert gold; VLM predictions as policy scores; engagement-style foils as negative controls; live product A/B out of scope. **Not demonstrated in this paper’s Results; no companion-track cells cited here.**
- **Stub assets.** Current public stubs are placeholders; redistribution requires cleared licenses.
- **Craft-domain ceiling.** Round-1 VLM **book/brand/tech accuracy** = 1.0 (all five cards) alongside human domain agreement = 1.0 (visual human $\approx$ 0.9875; only three Met-visual disagreements at 0.6667) is best read as an **easy-pair ceiling**—**not** general taste SOTA or universal mastery. Visual is where models separate ($\approx$ 0.84–0.91).
- **Post-hoc hard subset is a stress slice, not new gold.** Craft domains have **no VLM errors**, so brand/tech hard slices are low-confidence only (honest shortfall); visual includes Met soft-gold 013/015/016 + VLM-error ranking. Hard-subset overall accuracy $\approx$ 0.88–0.93 (visual $\approx$ 0.675–0.825; book/brand/tech still 1.0). Cite as disclosure, not as replacement leaderboard.
- **Met soft-gold / no fourth adjudicator.** Met visual 013/015/016 are 2/3 majority soft-gold; adjudication CSVs document votes; **no fourth rater**; primary gold keeps majority preference.

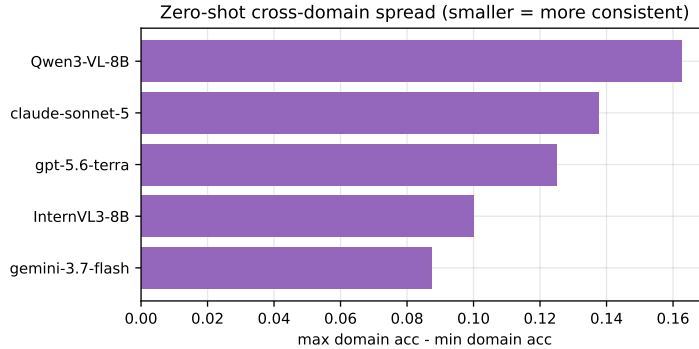


Figure 5: Zero-shot cross-domain spread per card (smaller is more consistent).

Artifact	Mechanism	Early detection
Position bias	A/B slot prior in construction or in the judging interface	report $P(\text{pref. in A})$; it should be ≈ 0.5
Lexical leakage	Predictive tokens in filenames / IDs (<code>good_</code> , <code>real_</code> , <code>v2_final</code>)	train a filename-only baseline; it should be $\approx$ chance
Difficulty ceiling	Consensus pairs an expert panel cannot make hard	report per-domain accuracy; a saturated domain is <i>unmeasured</i> , not solved

Table 11: Recurring artifacts in pairwise expert benchmarks, independent of domain or model.

- **Position/construction artifact.** Preferred/good items were often placed as `item_a`, so `always_item_a` ≈ 0.956 is a **construction artifact**, not a model ceiling; fair floors are `random_seed42` ≈ 0.552 / `hash_coin` = 0.5. The VLM run mitigates this via `eval_assignment_seed7` A/B randomization (`gold_side_A` ≈ 0.47); future rounds should randomize side at construction.
- **InternVL3-8B runtime substitute.** The card is `google/gemma-3-12b-it` after a free-tier stall—**not** InternVL3 weights. ECE and condition sensitivity are **not reported** for this run (predictions lack confidence; cue schema has triples but 0 conditioned rows in gold / assignment / predictions), though the shipped eval harness and `prompt_v2` / `prompt_cue_v1` now make both computable on a re-run; zero-shot cross-domain spread is reported (Table 10). Do **not** invent ECE or cue-sensitivity numbers from this run.
- **Companion observational proxy.** The companion paper’s observational q is a percentile rank of random `long_view_rate`, not expert aesthetic gold—do not import it as TasteBench success.

12 Ethics and Disclosure

- Use **licensed** or otherwise redistribution-permitted assets; record `license` per item.
- Do not show raters engagement statistics during gold labeling.
- Hash or anonymize rater identifiers in public releases; obtain consent consistent with institutional policy.
- Disclose any `data_origin=simulation` rows used for stress tests; never present simulation as human gold.

- Disclose Round-1 gold tier explicitly: `primary_human_expert` raters are biological humans (agent protocol = communication channel only). Do not narrate the superseded archival agent-panel file as current analysis gold.
- Dual-use: craft-discrimination tools can be abused for persuasive design or engagement hacking; we explicitly discourage optimizing TasteBench via engagement proxies and document negative controls to surface that failure mode.
- Cultural respect: circle tags are analytical labels, not pejoratives; guidelines forbid rewarding stereotypes as “taste.”

13 Conclusion

Expert-aligned taste benchmarks should reward craft judgment, not test-gaming. TasteBench pairs expert-only gold with a controlled protocol—a weak-label firewall, side randomization at construction, filename scrubbing, and hard filters—that makes the difference measurable. Applied to a Round-1 pilot (`primary_human_expert`; three biological human experts; $n = 250$), the protocol quantifies three artifacts—position bias (`always_item_a` ≈ 0.956), lexical leakage (`leak_prefer_good_token` ≈ 0.912), and a difficulty ceiling (craft saturated at 1.0 for humans and all five VLM cards; only visual separates, ≈ 0.84 – 0.91)—and shows that the two construction artifacts are *correctable* while the ceiling is a pair-difficulty limit. We release the eval harness, the side-randomization tool, and the prompt templates. The contribution is this controlled view: it tells practitioners which pitfalls to remove, and what must change—harder, contested, side-randomized pairs—before such a benchmark can support general taste claims. We deliberately decline the claims we cannot yet support (engagement, era/circle conditioning, calibration, general SOTA).

A Frozen defaults (cross-paper)

TasteBench (this paper): expert $n \geq 3$; percent agreement ≥ 0.60 (inclusive) primary gate; below $\rightarrow$ `difficulty=ambiguous` / `exclude_primary`; Krippendorff α appendix-only; hard-filter eval; weak $\rightarrow$ train/alignment only; test = expert (`bench_gold` / `label_aggregates` with `label_grain` / `preferred_id`); all bench tables carry `data_origin`; conditional pairs use `pair_group_id + cue_type` $\in \{\text{null, correct, misleading}\}$; H3 engagement-as-taste is a competing hypothesis, not gold. **Companion-only (cross-ref, not re-frozen here):** $\tau = 0.70$ (sweep 0.60/0.80); recency half-life 90d (sweep 30/90/180); H2 circle membership = `user_circles` (TasteBench may use item `circle_tags` as condition cues only).

B Toy eval (non-empirical)

We validated the eval harness on a 19-pair stub gold set: a chance-like predictor yields pairwise accuracy ≈ 0.42 and ECE ≈ 0.13 , while an oracle predictor yields 1.0. These figures confirm that the pipeline computes its metrics correctly; **they are not model performance**, and the stub set is a wiring fixture, not empirical data.

References

Pierre Bourdieu. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA, 1984. Translated by Richard Nice. Original French: *La Distinction*, 1979. Theory framing only, not a measurement identity. No DOI in draft.

- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.
- Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. doi: 10.48550/arXiv.1706.03741.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pages 1321–1330, 2017.
- Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM ’17)*, pages 781–789. ACM, 2017. doi: 10.1145/3018661.3018699.
- Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charles Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision (ECCV)*, 2016.
- Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE, 4 edition, 2018. doi: 10.4135/9781071878781. α reliability; appendix-only in this paper.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint*, 2023.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multimodal model an all-around player? *arXiv preprint*, 2023.
- Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2408–2415, 2012. doi: 10.1109/CVPR.2012.6247954.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022. doi: 10.48550/arXiv.2203.02155.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, 2023.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024.