IDENTIFIABLE LATENT RUBRICS: FROM PREFERENCE FIT TO SEMANTIC VALIDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can reusable evaluation criteria be learned without updating a language-model backbone? We study parameter-efficient semantic grounding on frozen representations: our Qwen3-32B main study trains only a 15,363-parameter rubric head, with no backbone updates. The central challenge is identification, not parameter scale. Preference likelihood is invariant to changes of latent basis; exact named interventions identify directions up to positive scales under a restricted linear model. A synthetic positive control raises axis recovery from 0.730 to 0.996 with one intervention per criterion, while preference accuracy changes little. On real text, a human-audited filter passes 19/20 fresh blind pairs, but locally valid edits alone do not ensure transfer: Silver-200 obtains 61.40% TEST rubric composition versus 72.57% for DEV-aligned preference-only training. A subsequent 10,240-parameter semantic-ranking head, using additional human labels and retaining the frozen backbone, reaches 72.67% DEV composition: 3.81 points above the natural-contrast intervention variant, but not a confirmed improvement over aligned BT. This demonstrates a lightweight route to improving development-set semantic metrics, not established TEST superiority. Our framework combines restricted identification theory, auditable supervision, and controlled diagnostics to distinguish parameter-efficient learning from the stronger requirement of reusable semantic axes.

1 INTRODUCTION

Learning from human preferences supports instruction-following alignment (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023), but a comparison reveals which response is preferred, not which combination of correctness, coherence, or other properties explains that preference. A reusable evaluation rubric requires named coordinates whose meanings survive changes of prompt, training seed, and evaluation rule.

Our practical setting keeps the pretrained language model frozen and learns only a compact interface over its cached representations. This allows criteria and objectives to be revised without backpropagation through the backbone or repeated feature extraction for unchanged responses. The main three-coordinate head has 15,363 trainable parameters; a separate semantic-ranking variant uses only 10,240. The opportunity is to improve semantic usefulness by changing a small readout, rather than the underlying language model. Parameter efficiency, however, does not by itself establish the meaning of that readout.

Structured rewards and latent preference factors make such representations explicit (Gunjal et al., 2026; Srivastava et al., 2026; Gong et al., 2025; Liu et al., 2026). Yet exposing multiple coordinates does not establish their interpretation. For $r = w^\top z$, any invertible A yields the same reward through $z' = Az$ and $w' = A^{-\top}w$. A single global readout need not even identify the full latent subspace. This distinction between usefulness and identifiability connects reward modeling to representation learning (Locatello et al., 2019; Khemakhem et al., 2020).

Named interventions supply a potential semantic anchor: an edit intended to change correctness should move its coordinate while preserving the others. Under exact linear assumptions, criterion names fix permutations and signed changes fix axis directions. This connects our formulation to weakly supervised and causal representation learning (Locatello et al., 2020; Brehmer et al., 2022; Schölkopf et al., 2021). Figure 1 contrasts preference fitting with this intervention-and-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.

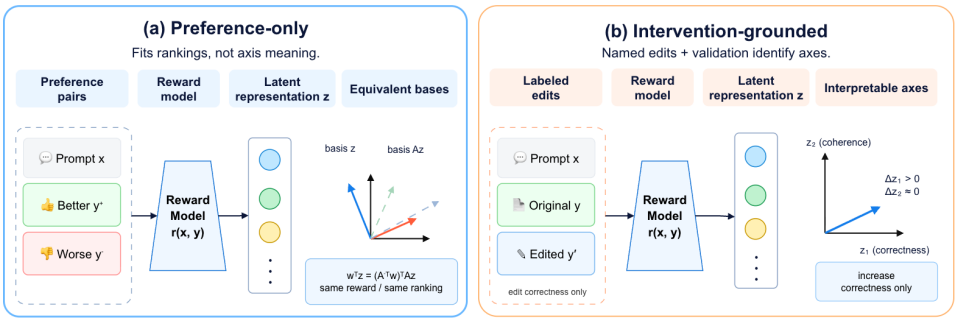

Figure 1: **From preference fitting to intervention-grounded semantic identification.** (a) **Preference-only:** pairwise preference supervision can fit a reward model that predicts rankings, but the latent coordinates remain non-identifiable: different latent bases can induce the same reward and the same rankings. (b) **Intervention-grounded:** labeled edits provide directional supervision for specific attributes, and independent validation is used to test whether the learned latent axes retain the intended semantic meanings.

approach; its semantic interpretation is a hypothesis to test, not an automatic consequence of adding a loss.

Our empirical contribution is to make those assumptions observable. One exact synthetic anchor per criterion raises axis recovery from 0.730 to 0.996. Human audits then distinguish selective natural-language edits from transformations that unintentionally change other attributes, and motivate a balanced correctness/coherence study. Importantly, better local edit selection does not guarantee transfer: Silver-200 achieves 61.40% TEST composition versus 72.57% for DEV-aligned BT. Targeted diagnostics show that finite, template-concentrated edits leave substantial semantic freedom; additional human supervision provides exploratory recovery, not independent confirmation.

We contribute a frozen-backbone framework for learning compact rubric heads, a precise account of preference equivalence and restricted axis identification, and a human-audited protocol that makes semantic supervision testable. Controlled experiments establish recovery under the theory's assumptions; real-text diagnostics and head-only recovery experiments identify which supervision and transfer requirements remain unresolved. The resulting contribution is both practical—a small, replaceable semantic interface—and methodological: measured improvement is separated from identification and independent generalization.

2 Related Work

Preference learning and structured rewards. Bradley–Terry comparisons (Bradley & Terry, 1952) underlie preference reward learning and instruction alignment (Christiano et al., 2017; Ouyang et al., 2022); DPO directly optimizes preferences without an explicit reward-training stage (Rafailov et al., 2023). Rubrics as Rewards supplies instance-specific criteria (Gunjal et al., 2026), while CROME uses causal-rubric augmentation for robustness (Srivastava et al., 2026). Latent Preference Coding and SparseRM expose discrete or sparse preference factors (Gong et al., 2025; Liu et al., 2026). We ask whether latent coordinates retain predetermined semantic meanings, not merely whether structured representations predict preference.

Identification and semantic supervision. Unsupervised disentanglement requires assumptions beyond observational fit (Locatello et al., 2019). Auxiliary variables, conditional models, paired weak supervision, and intervention structure offer identification routes (Hyvärinen et al., 2019; Khemakhem et al., 2020; Locatello et al., 2020; Brehmer et al., 2022; Schölkopf et al., 2021). Our propositions specialize these ideas to linear reward coordinates, without claiming a general impossibility theorem or recovering a causal graph. Concept bottlenecks illustrate explicit semantic supervision (Koh et al., 2020); our dense attribute reference uses such extra information and is not budget-matched.

Counterfactual validity and evaluation. RATE studies reward explanations with imperfect counterfactual rewrites (Reber et al., 2025); we instead test coordinate semantics and compositional reuse. HelpSteer2 and its preference extension provide complementary attribute and preference labels (Wang et al., 2024; 2025). Known judge biases (Zheng et al., 2023) motivate randomized A/B presentation, hidden intervention metadata, and human validation rather than model self-certification.

3 Setup and Identification

Preference supervision. For prompt x and responses y^+, y^-, the Bradley–Terry model (Bradley & Terry, 1952) uses

$$P(y^+ \succ y^- \mid x) = \sigma(r(x, y^+) - r(x, y^-)), \qquad \mathcal{L}_{\text{pref}} = -\log P(y^+ \succ y^- \mid x). \tag{1}$$

A frozen representation $h(x, y) \in \mathbb{R}^d$ is projected to $z(x, y) = Ch(x, y)$ and aggregated as $r(x, y) = w^\top z(x, y)$. Four human attributes—correctness, coherence, complexity, and verbosity—provide semantic evaluation targets. Helpfulness is treated as holistic rather than an additional atomic axis. The attribute scores are not training targets for the preference-only or intervention models; dense attribute supervision is explicitly separated.

What we mean by a semantic rubric. Formal identification concerns uniqueness within a specified model class, up to stated transformations. Our empirical diagnostics are weaker: human-attribute alignment, cross-seed coordinate stability, selective responses to changes, and composition into held-out weighted rubrics. No single diagnostic proves identification. In particular, marginal correlations can be confounded by dependent human attributes, and generated-edit selectivity may reflect an edit template rather than the intended concept.

3.1 What Preferences Leave Unidentified

Proposition 1 (Preference-preserving reparameterization)**.** *For any invertible $A \in \mathbb{R}^{K\times K}$, $z' = Az$ and $w' = A^{-\top}w$ induce the same scalar reward and the same Bradley–Terry probabilities as (z, w).*

The proof is the identity $(A^{-\top}w)^\top Az = w^\top z$. This concerns likelihood, not the invariance of a regularized optimizer. Orthogonality may restrict the admissible transformations but does not provide semantic names. Moreover, with one global linear readout the reward depends only on $q = C^\top w$. Different row spaces of C can induce the same q (Appendix A); recovery of a complete subspace is not guaranteed.

3.2 A Restricted Sufficient Condition

Let $c(x, y) \in \mathbb{R}^K$ denote semantic coordinates. A named intervention is a within-prompt pair with $\Delta c = te_k$, $t \neq 0$, and known identity k and direction $s = \text{sign}(t)$.

Proposition 2 (Exact named-axis grounding)**.** *Suppose $z(x, y) = Mc(x, y) + b(x)$ for a single invertible matrix M shared across prompts. For every k, assume at least one named one-factor intervention exists. If each such pair satisfies $\Delta z = \gamma e_k$ with $s\gamma > 0$, then M is positive diagonal.*

Indeed, $tMe_k = \gamma e_k$ gives $Me_k = (\gamma/t)e_k$ with $\gamma/t > 0$. Coverage of every criterion constrains every column. Thus names fix permutations and directions fix signs, while positive scales and prompt-dependent offsets remain free.

This is an idealized sufficient condition, not an optimization guarantee or a universal necessity result. Other supervision or distributional assumptions can identify representations (Hyvärinen et al., 2019; Khemakhem et al., 2020; Koh et al., 2020). Partial anchoring also leaves additional ambiguity. With named variables c and unobserved nuisance variables n, pure interventions on c can constrain a mixing matrix only to

$$M = \begin{bmatrix} D & B \\ 0 & R \end{bmatrix}, \tag{2}$$

where D is positive diagonal, but B may mix nuisance variation into named outputs. Our residual-augmented heads therefore require independent semantic tests; pure named interventions alone do not establish globally nuisance-free coordinates.

Operational intervention validity. For human signed changes $\Delta a_j \in \{-2, -1, 0, 1, 2\}$, a target hit requires $s\Delta a_k \geq 1$. The historical four-attribute one-factor criterion additionally requires $\Delta a_j = 0$ for every other measured dimension. The fresh V2 audit strengthens this criterion by also requiring no flagged other meaningful change. We retain the historical definition when reporting the earlier audit and follow-up. Human zeros are finite-resolution judgments, not proof that every latent factor is invariant. Correctness is task-relative: an edit must improve or degrade satisfaction of the task's requirements, not merely change a factual statement in isolation.

4 Learning and Validating Latent Rubrics

We freeze the pretrained backbone, cache its response representations, and learn a small rubric head from preference rankings and named edits (Figure 2). Only the head receives gradients. The diagram shows the two named axes, correctness and coherence; the implemented head also includes a separate residual.

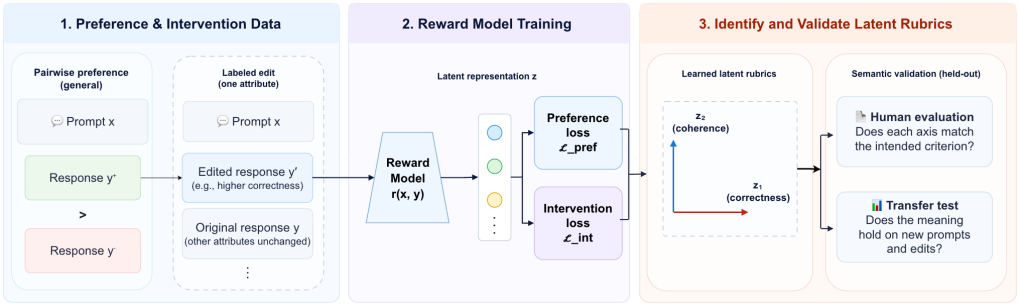

Figure 2: **Overview of our framework.** We combine pairwise preference data with labeled edits to train a reward model with intervention objectives. In our implementation, the backbone is frozen and only a compact rubric head is trained. The learned latent space is then validated through independent human evaluation and transfer tests to determine whether each dimension corresponds to a meaningful and generalizable criterion.

4.1 Directional Anchoring

For frozen features $h_\theta(x, y) \in \mathbb{R}^{5120}$, we optimize $C \in \mathbb{R}^{3\times 5120}$ and $w \in \mathbb{R}^3$, while θ remains fixed. Thus the main study updates $3 \cdot 5120 + 3 = 15{,}363$ parameters, not the 32B backbone. No adapter is inserted into the completed main model. A separate exploratory ranking head updates two semantic rows only (10,240 parameters) and keeps an existing BT reward fixed. Appendix B derives the objectives, gradients, and parameter accounting.

The preference-only model, Latent-BT, fits $\mathcal{L}_{\text{pref}}$ through a linear head. For edit $(x, y, \tilde{y}, k, s)$, write $\Delta z = z(x, \tilde{y}) - z(x, y)$. The interventional model ILR adds

$$\mathcal{L}_{\text{target}} = -\log \sigma(s\Delta z_k/\tau), \tag{3}$$

$$\mathcal{L}_{\text{leak}} = \frac{1}{K-1}\sum_{j\neq k} \text{Huber}(\Delta z_j), \qquad \mathcal{L}_{\text{res}} = \text{Huber}(\Delta z_{\text{res}}), \tag{4}$$

$$\mathcal{L} = \mathcal{L}_{\text{pref}} + \lambda_t \mathcal{L}_{\text{target}} + \lambda_l \mathcal{L}_{\text{leak}} + \lambda_r \mathcal{L}_{\text{res}}. \tag{5}$$

K counts named axes, excluding the optional residual. Synthetic models use four axes without a residual; historical text models use four plus a residual; the main follow-up uses $[z_C, z_{Coh}, z_{\text{res}}]$. The losses encourage selectivity but neither certify edit semantics nor eliminate the nuisance shear in Eq. (2).

4.2 Human-Audited Supervision

An initial 80-pair audit diagnoses criterion-specific construction failures. A fresh 20-pair blind packet then tests filter V2, with four high-ranked and one boundary example in each criterion–direction cell. A/B order is randomized; criterion, direction, score, generator, and original/edited

identities are hidden. A private key restores signed human judgments. After examining one failure, V2.1 adds a task-relative correctness guard; the V2 results remain frozen rather than becoming independent validation of that revision.

The filter uses direct offline GPT-6 review, not an independent model jury. Scores 95/85/75 are ordinal selection judgments, not calibrated probabilities or training targets. Every selected family receives full-context review. Candidate families and prompts in either human audit are excluded consistently from intervention training, preference training, dense attribute training, and TRAIN normalization. Explanations never enter training. Exact prompt–response identities link cached features to edit endpoints; Appendices D and E give the audit and leakage checks.

5 EXPERIMENTS

We test the anchoring mechanism under controlled assumptions, audit its natural-language supervision, and measure semantic transfer. The completed main TEST study remains frozen; subsequent development-set searches are explicitly exploratory. Boldface highlights focal results and parameter budgets, not statistical significance.

5.1 CONTROLLED IDENTIFICATION AND EVALUATION PROTOCOL

Real-text experiments use HelpSteer2 and its preference extension (Wang et al., 2024; 2025), with 10,669 prompts and 21,360 responses. TEST contains 518 prompts, 1,038 responses, and 352 non-tie preference pairs. The main representation is frozen Qwen3-32B (Yang et al., 2025), block 48, response-mean pooling, with linear heads and seeds 0–4. Preference-only, intervention-trained, and dense human-attribute models share preference data; the dense reference receives additional semantic labels.

We distinguish preference accuracy, axis purity, natural target effect, and rubric composition. Purity measures the target-minus-largest-off-target absolute Spearman correlation. Target effect is the signed coordinate change divided by its TRAIN SD; natural pairs isolate a change in observed human ratings and remain observational, not randomized interventions. Composition tests within-prompt rankings under weighted human rubrics. The historical four-axis study uses 64 weights and TEST standardization; the main two-axis study uses 21 nonnegative weights and allowed-TRAIN scales for both latent and human scores. BT axes are permuted and signed on DEV; named axes are not rematched. These protocols are not numerically comparable across studies. Reported SDs describe five fitted seeds; new 95% intervals use 10,000 paired prompt-cluster resamples with those fits fixed.

The synthetic positive control mixes four Gaussian factors into 64 dimensions with observation noise SD 0.1. Fifty reward contexts provide richer subspace information than one global readout; TRAIN/DEV/TEST contain 20,000/5,000/5,000 preference pairs. Eight nested intervention budgets reveal criterion names and directions, not magnitudes. As Figure 3 shows, preference-only learning already recovers the subspace (0.9937) but not its basis (0.7300). One anchor per criterion raises axis recovery to 0.9963 and composition from 77.2% to 97.5%, while preference accuracy stays near 82.4%. At $N = 25$, the preference difference is .00244 with interval $[-.00252, .00740]$, not an equivalence finding.

Table 1 tests whether generic regularization explains this recovery. Adding orthogonality without anchors leaves axis recovery at .728 and composition at 76.97%, compared with .730 and 77.18% for preference-only learning. Named anchors reach 99.39% composition with the same 456-parameter head. Imperfect interventions degrade this benefit gradually: adding independent Gaussian changes to the non-target factors gives 98.61%, 97.48%, 95.35%, and 90.39% composition at noise SDs .05, .1, .2, and .5. At the largest tested noise, recovery remains .953 but preference accuracy falls to 80.98%; robustness is therefore partial, not cost-free. These five-seed controls support semantic anchoring beyond generic orthogonality within this linear data-generating process. Direction matters separately: the single-seed sign-flipped control retains .9998 absolute recovery but has signed target effect $-.9974$. High unsigned recovery alone can conceal reversed meanings.

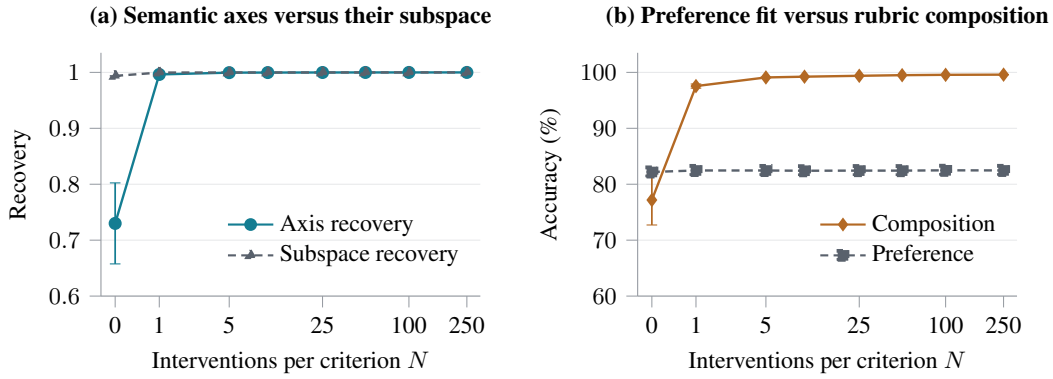

Figure 3: Exact anchors resolve a semantic ambiguity that preference accuracy barely exposes. Means and ± 1 sample SD over five seeds at eight completed budgets; the horizontal position is logarithmic in $N + 1$ and labels show N. With no anchors, the subspace is recovered but individual axes are not. One anchor per criterion nearly closes the axis-recovery gap and improves composition, while preference accuracy changes little. Axis recovery uses absolute factor correlations; subspace recovery uses canonical correlations (Appendix F).

Table 1: Synthetic mechanism and noise controls, mean $\pm$ sample SD over five seeds. All models train the same 456 parameters. Anchored rows use $N = 25$ per criterion; σ_{off} is non-target semantic noise SD, with target magnitude sampled from $[.5, 1.5]$. Unnamed-axis recovery permits optimal matching; named axes retain their order.

Training condition	Axis recovery	Preference (%)	Composition (%)
Preference only	$.7300 \pm .0724$	$82.19 \pm .08$	77.18 ± 4.47
Preference + orthogonality	$.7284 \pm .0657$	$82.23 \pm .09$	76.97 ± 4.17
Named anchors, $\sigma_{\text{off}} = 0$	$\mathbf{.9998 \pm .00002}$	$82.44 \pm .06$	$\mathbf{99.39 \pm .02}$
Named anchors, $\sigma_{\text{off}} = .05$	$.9990 \pm .0001$	$82.42 \pm .14$	$98.61 \pm .08$
Named anchors, $\sigma_{\text{off}} = .1$	$.9967 \pm .0005$	$82.39 \pm .26$	$97.48 \pm .20$
Named anchors, $\sigma_{\text{off}} = .2$	$.9887 \pm .0018$	$81.94 \pm .29$	$95.35 \pm .43$
Named anchors, $\sigma_{\text{off}} = .5$	$\mathbf{.9529 \pm .0095}$	$80.98 \pm .55$	$\mathbf{90.39 \pm 1.06}$

5.2 Human Validation and Real-Text Transfer

The initial four-axis text study illustrates why supervision needs validation. GLM-5.3 edits checked by a local Qwen3-32B judge yield stronger generated-edit target effects (0.0816 to 1.1129), but lower natural effects (0.2408 to 0.1220) and composition (60.6% to 56.7%). The paired composition difference is -3.95 points, with interval $[-8.08, -.28]$. This directionally imbalanced pool lacks coherence-positive anchors. The 14B replication and ten 32B layer/pooling checks have the same descriptive composition ordering; full results and historical metric caveats remain in Appendix F.

An 80-pair human audit, balanced across criteria and directions, identifies selective transformations for the two-axis follow-up. The enriched diagnostic sample contains 56 borderline/single-judge-clean and 24 likely-impure examples, not a random precision survey. Correctness/coherence satisfy the historical four-attribute one-factor rule in 17/20 and 18/20 cases, versus 3/20 for complexity and 2/20 for verbosity (Figure 4). Verbosity changes its target in all 20 cases but also changes coherence in 18. Adding the later no-other-meaningful-change requirement reduces coherence to 17/20. These results diagnose constructions, not intrinsic inseparability of concepts.

After an initial filter gate fails (6/7 accepted examples pass, with insufficient coverage), a fresh V2 blind test yields 19/20 passes under the stricter rule: 16/16 high-ranked and 3/4 boundary. Wilson 95% intervals are 76.4–99.1%, 80.6–100%, and 30.1–95.4% (Wilson, 1927). The result supports precision-oriented collection, not perfect precision or a controlled improvement over the earlier gate. One task-relative correctness failure motivates V2.1 without independently validating it. Its Silver-200 set contains 200 directed rows from 100 distinct prompts and edit families, balanced

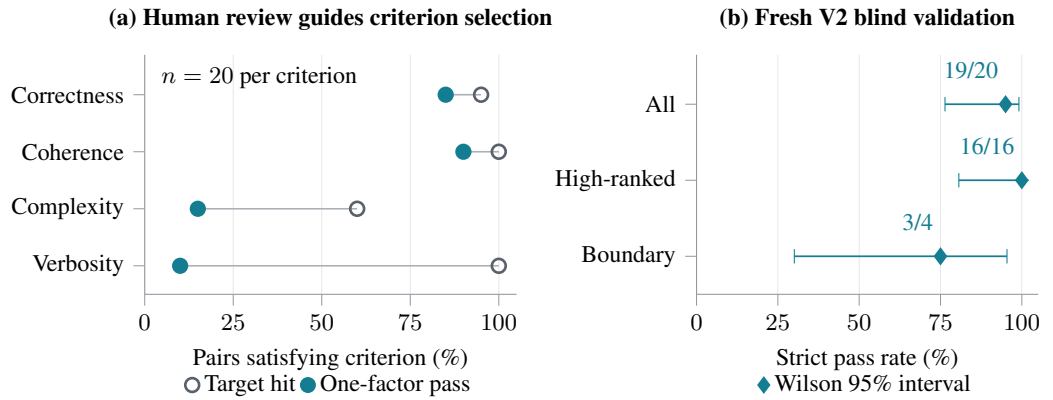

Figure 4: Human evidence diagnoses supervision and quantifies the remaining uncertainty. Left: each gap separates an intended target change from a four-attribute one-factor pass in the enriched 80-pair audit; the historical pass rule does not include the additional-change flag. Right: V2 uses the stricter rule on 20 fresh prompts, with four high-ranked and one boundary pair per criterion–direction cell. All is the pooled sample, not an independent stratum. The high-ranked 16/16 outcome is encouraging but its lower confidence bound is 80.6%; none of these estimates validates the subsequently revised V2.1 rule.

across correctness/coherence directions; reverse rows are not independent observations. Both human audits are excluded. Construction details and the initial TinyHuman diagnostic are retained in Appendices E and F.

Twenty completed fits compare aligned BT, Silver-200, nested Silver-100, and a dense attribute reference. Table 2 retains the main negative evidence: Silver-200 composition is 11.17 points below BT, with paired interval $[-14.90, -7.44]$; strict natural target effect differs by $-.165$ with interval $[-.313, -.028]$. Natural support is 111 pairs, predominantly correctness (94 versus 17 coherence), rather than an equal-criterion average. Preference accuracy is close in mean, but the difference interval $[-2.30, 1.88]$ points does not establish equivalence. Doubling Silver does not improve mean composition, and dense labels are not a guaranteed upper bound. Local edit credibility therefore does not establish transferable axes.

Table 2: Main two-axis TEST results, mean $\pm$ sample SD over five fits. BT axes are DEV-aligned; dense training receives extra human attributes. Complete metrics and paired intervals are in Appendix H.

Model	Preference (%)	Composition (%)	Natural TE
Aligned BT	$74.26 \pm .65$	$72.57 \pm .76$	$.272 \pm .026$
Silver-100	73.98 ± 3.02	62.78 ± 6.45	$.123 \pm .092$
Silver-200	74.03 ± 3.17	61.40 ± 8.58	$.107 \pm .106$
Dense attribute	$74.49 \pm .74$	66.63 ± 1.34	$.181 \pm .044$

5.3 Head-Only Recovery and the Remaining Identification Gap

The Silver models fit both criteria's training directions perfectly. Same-pipeline endpoint re-extraction and retraining barely change DEV composition (59.74% to 59.72%), narrowing a feature-provenance explanation. Holding out 20 intervention families retains perfect direction accuracy but only 58.38% composition; their prompts remain in preference training, so this is not fully unseen-prompt generalization. A rule based on the negative sum of integer tokens predicts every correctness training direction while leaving coherence edits unchanged. This demonstrates a compatible shortcut, not the model's actual mechanism.

More generally, 100 independent feature differences have rank 100 in 5,120 dimensions. Nullspace perturbations with compensating residual changes preserve all scalar rewards and observed intervention deltas while changing DEV composition. Their regularized norms differ; this is observational underconstraint, not equivalence of the full optimization objective. These diagnostics connect the real-text failure to the theorem's global structural assumptions rather than merely insufficient optimization.

Semantic recovery can be pursued through the small head alone. Table 3 brings the post-study comparisons onto common prompt exclusions and TRAIN scales, separate from the original TEST study. Replacing Silver's 100 synthetic-edit families with 100 natural human-contrast families raises mean DEV composition from 64.45% to 66.85%; expanding to 400 natural families reaches 68.86%. These are descriptive comparisons, not established direction-accuracy gains on the 100 excluded natural prompts. Full-attribute joint training reaches only 65.60%, so access to more labels does not alone determine the outcome.

A separate semantic-ranking objective reaches $72.67 \pm .11\%$ with just 10,240 newly trained parameters, versus 15,363 for the joint heads. Relative to HumanDelta-800, its gain is $+3.81$ points with paired interval $[1.51, 6.07]$; relative to aligned BT, the $+.73$-point interval $[-1.27, 2.72]$ includes zero. All three tested regularization settings reach 70.91–72.67% composition; the selected setting minimizes DEV ranking loss, not composition. This supports a lightweight route to development-set recovery, but changes supervision and selection as well as the objective: it does not isolate gradient interference or establish TEST superiority. Preference is preserved structurally by retaining the BT reward as a residual with zero semantic readout weights, not by showing a faithful semantic reward decomposition. Broader frozen-feature screening peaks at 72.74%; interrupted LoRA runs have no final composition. Appendices H and I retain the complete search and limitations.

Table 3: Frozen-backbone head comparisons on DEV only, mean $\pm$ sample SD over five fits. All rows share the later prompt exclusions and TRAIN normalization. Supervision and selection differ; this is not a budget-matched ablation or independent TEST confirmation. "Full TRAIN" uses the available human attribute ratings; ranking omits per-axis ties.

Head / objective	Semantic supervision	Trained parameters	Composition (%)
Aligned BT	DEV alignment only	15,363	$71.94 \pm .56$
Silver-200, joint	100 edit families	15,363	64.45 ± 1.90
HumanDelta-200, joint	100 natural families	15,363	66.85 ± 2.01
HumanDelta-800, joint	400 natural families	15,363	68.86 ± 1.34
Full-attribute, joint	Full TRAIN	15,363	65.60 ± 1.90
Separate semantic ranking	Full TRAIN	**10,240**[†]	$\mathbf{72.67 \pm .11}$

[†]Newly optimized semantic parameters only; the existing BT reward is fixed and attached afterward. Paired 95% intervals: ranking minus HumanDelta-800, $\mathbf{+3.81}$ points $[1.51, 6.07]$; ranking minus BT, $+.73$ points $[-1.27, 2.72]$. These condition on DEV-selected fits.

6 Discussion and Limitations

Frozen-backbone learning makes rubric revision a small-head problem: cached features support repeated optimization without changing the language model. The controlled recovery and exploratory semantic-ranking results show that this interface can improve semantic metrics. We establish the trainable-parameter budget, not an end-to-end speedup: feature extraction still requires backbone inference, and no matched full-fine-tuning cost benchmark is available. Scientifically, three requirements remain distinct: *validity* of an edit, *coverage* of semantic variation, and *learning* coordinates that transfer. The exact-anchor control supports the restricted theory; human audits test the supervision; the Silver result shows why valid local edits are insufficient. This proxy-versus-target distinction also motivates reward-overoptimization research (Gao et al., 2023), although we do not optimize policies.

The evidence has clear limits. Human audits are small and selectively sampled, with no independent multi-annotator reliability estimate or fresh V2.1 precision test. Silver covers only 100 narrow-template families; natural response pairs are observational. The historical verifier shares the feature

model's family. Original four-axis TEST normalization differs from the TRAIN-scaled main comparison. New intervals condition on fixed fits and frozen mappings, are unadjusted across outcomes, and do not cover repeated DEV selection. The separated head uses extra labels and preserves preference through a distinct BT residual; its result neither isolates reduced gradient interference nor explains reward through semantic coefficients. A frozen candidate, broader cross-construction holdouts, and independent evaluation are required before claiming a reusable real-text rubric.

7 Conclusion

Reusable rubrics can be studied through a compact trainable head on a frozen backbone. Exact named interventions establish identification under restricted assumptions; controlled recovery and head-only DEV improvements demonstrate the potential of this parameter-efficient interface. Real-text transfer remains the stronger test: validate edits, broaden their semantic coverage, and evaluate a frozen candidate independently. Small parameter counts, preference fit, and local edit selectivity are useful properties, but none alone establishes reusable semantic axes.

Impact Statement

Named reward coordinates may make evaluation more transparent, but incorrect semantic names can also create unwarranted confidence in automated judgments. Our results caution against deploying intervention-sensitive coordinates as trusted human rubrics without independent validation. The released edits include deliberately degraded text and are intended for controlled research. We do not demonstrate a deployed safety benefit or improved policy alignment.

Ethics Statement

The experiments use benchmark model responses and constructed edits as research stimuli, not a deployment-facing decision system. Some prompts and deliberately degraded responses may contain sensitive, misleading, or otherwise undesirable content; inclusion is for evaluation, not endorsement. Human annotations and model reviews are distinguished in the release. Dataset licensing, annotation consent, and the access scope of released audit records should be checked by the authors before dissemination; this manuscript does not assert an ethics approval or consent procedure absent from the available records.

Reproducibility Statement

The archive preserves configurations, seeds, predictions, and metrics for the synthetic and historical real studies, the initial failed gate, and the completed TinyHuman branch. The latest snapshot adds 20 completed two-axis fits, frozen mappings, TRAIN scales, TEST predictions, and 10,000 paired prompt-bootstrap draws. Separate artifacts record same-pipeline endpoint diagnostics, natural-contrast and separated-head studies, corrected ensemble scores, and an interrupted LoRA branch with no final result. The original Silver preparation manifest describes an earlier state and is not used to infer current training status. All 1,395 payload files in the update manifest passed SHA256 verification; figure data are recomputed from saved fits. The evidence map, aggregate chart data, citation ledger, and build instructions accompany the manuscript. Rewriting and verification did not rerun training or add new human labels.

AI Use Statement

AI tools assisted with drafting, conceptual figure preparation, code execution, candidate generation, and model-based screening. V2/V2.1 scores come from direct GPT-6 review and are model judgments, not human ground truth or independent multi-model consensus. Human-supplied annotations determine the reported audit outcomes. The authors remain responsible for reviewing the scientific claims, cited sources, disclosure requirements, and final manuscript.

REFERENCES

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. URL https://academic.oup.com/biomet/article-abstract/39/3-4/324/326091.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38319–38331, 2022. doi: 10.52202/068431-2776. URL https://papers.neurips.cc/paper_files/paper/2022/hash/fa567e2b2c870f8f09a87b6e73370869-Abstract-Conference.html.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Zhuocheng Gong, Jian Guan, Wei Wu, Huishuai Zhang, and Dongyan Zhao. Latent preference coding: Aligning large language models via discrete latent codes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19897–19911. PMLR, 2025. URL https://proceedings.mlr.press/v267/gong25a.html.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cfd7bee7a651ee9af525098ef67a9e45-Abstract-Conference.html.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Dengcan Liu, Jiahao Li, Zheren Fu, Yi Tu, Jiajun Li, Zhendong Mao, and Yongdong Zhang. SparseRM: A lightweight preference modeling with sparse autoencoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32051–32059, 2026. doi: 10.1609/aaai.v40i38.40476. URL https://ojs.aaai.org/index.php/AAAI/article/view/40476.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359. PMLR, 2020. URL https://proceedings.mlr.press/v119/locatello20a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

David Reber, Sean M. Richardson, Todd Nief, Cristina Garbacea, and Victor Veitch. RATE: Causal explainability of reward models with imperfect counterfactuals. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51341–51368. PMLR, 2025. URL https://proceedings.mlr.press/v267/reber25a.html.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL https://doi.org/10.1109/JPROC.2021.3058954.

Pragya Srivastava, Harman Singh, Rahul Madhavan, Gandharv Patil, Sravanti Addepalli, Arun Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuveer, Karthikeyan Shanmugam, and Doina Precup. Robust reward modeling via causal rubrics. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fa0a1013e171c1be6121c9bc9fb6589f-Abstract-Conference.html.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer 2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0047. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/02fd91a387a6a5a5751e81b58a75af90-Abstract-Datasets_and_Benchmarks_Track.html.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. HelpSteer2-Preference: Complementing ratings with preferences. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8e237ec6d3bc86f6d4967e9c7eef37ff-Abstract-Conference.html.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL https://www.jstor.org/stable/2276774.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger

Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Identification: Proofs and Scope

A.1 Observation Model and Equivalence Class

The representation $h_\theta(x, y) \in \mathbb{R}^d$ is fixed throughout head training. Write $z = Ch_\theta$ and $r = w^\top z$, with $C \in \mathbb{R}^{L\times d}$ and $w \in \mathbb{R}^L$. Here $L = K$ in the fully named theoretical model and $L = K + 1$ for the empirical head with one residual. A preference pair supplies a label $b \in \{0, 1\}$ for responses y_a, y_b. Its likelihood depends on only one scalar,

$$u(x, y_a, y_b) = w^\top C\big(h_\theta(x, y_a) - h_\theta(x, y_b)\big), \qquad P(b \mid u) = \sigma(u)^b[1 - \sigma(u)]^{1-b}. \tag{6}$$

Thus observing accurate preference predictions does not directly reveal the L latent coordinates. A prompt-dependent reward offset $a(x)$ also cancels from every within-prompt logit. This is why the identification result below concerns coordinate directions rather than absolute levels.

Full proof of Proposition 1. Choose any $A \in \mathrm{GL}(L)$ and define $C' = AC$, $w' = A^{-\top}w$. Then for every feature vector, not just the observed sample,

$$z' = ACh_\theta = Az, \tag{7}$$

$$w'^\top z' = (A^{-\top}w)^\top Az = w^\top A^{-1}Az = w^\top z, \tag{8}$$

$$u' = w'^\top C'(h_a - h_b) = u. \tag{9}$$

Substitution into Eq. (6) gives equal likelihood for either value of b and hence equal empirical preference loss on every dataset. The transformation can rotate, rescale, or shear the latent basis. It need not preserve coordinate names. This argument is about the likelihood: weight decay or another regularizer may prefer particular representatives without making them semantically identified.

The effective reward direction is $q = C^\top w \in \mathbb{R}^d$. Even q is only partially determined on a finite sample. If H_Δ contains all observed preference differences as columns, then $q' = q + v$ produces the same observed logits whenever $H_\Delta^\top v = 0$. This finite-design ambiguity is separate from the factorization ambiguity $(C, w) \mapsto (AC, A^{-\top}w)$, which preserves rewards for all h.

A.2 Why One Reward Does Not Identify a Latent Subspace

With $K = 2, d = 3, w = (1, 0)^\top$, let

$$C_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Both rewards equal $h_1(x, y)$, but the first row space is $\mathrm{span}\{e_1, e_2\}$ and the second is $\mathrm{span}\{e_1, e_3\}$. No invertible change of a two-dimensional basis turns one row space into the other. Preference equivalence can therefore be broader than a rotation within an otherwise identified subspace.

Multiple reward contexts provide different information. If context readouts form $W = [w_1, \ldots, w_T] \in \mathbb{R}^{K\times T}$, their effective directions are $Q = C^\top W$. When C has row rank K and W has rank K, $\mathrm{col}(Q) = \mathrm{col}(C^\top)$. This linear-algebra observation explains why the synthetic design uses 50 diverse contexts. It does not establish that finite noisy preferences recover Q, or that the real-text setting with one global readout identifies a complete subspace. The synthetic subspace result is accordingly a positive control under a richer observation model.

Orthogonality does not assign names either. If $CC^\top = I$ and A is orthogonal, then $(AC)(AC)^\top = I$ while the preference likelihood is unchanged. An orthogonality constraint removes some scale and shear freedom, but leaves rotations and permutations of unnamed axes.

A.3 Exact Named Interventions: Complete Column-Wise Proof

The restricted structural assumption is $z(x, y) = Mc(x, y) + b(x)$ for one invertible $M \in \mathbb{R}^{K\times K}$ shared across prompts. It is stronger than assuming only that a large frozen representation contains some semantic information. For each k, suppose there is a same-prompt pair satisfying

$$c(x, \tilde{y}) - c(x, y) = t_k e_k, \qquad t_k \neq 0, \qquad s_k = \mathrm{sign}(t_k). \tag{10}$$

Subtracting the two representations eliminates $b(x)$:

$$\Delta z = M\Delta c = t_k M e_k. \tag{11}$$

Exact latent selectivity additionally gives $\Delta z = \gamma_k e_k$ and $s_k\gamma_k > 0$. Comparing each component yields

$$M_{jk} = 0 \quad (j \neq k), \qquad M_{kk} = \frac{\gamma_k}{t_k} > 0. \tag{12}$$

Repeating this argument for every k constrains every column, so

$$M = \mathrm{diag}(d_1, \ldots, d_K), \qquad d_k > 0. \tag{13}$$

Conversely, any such positive diagonal matrix satisfies the required selectivity and direction for every pure intervention. The remaining scale freedom is genuine: signs alone cannot distinguish $d_k = 1$ from $d_k = 10$. The prompt offset is also unconstrained by within-prompt changes. Known names remove permutations; known directions remove sign flips; neither supplies metric units.

Coverage and shared structure are essential. If a criterion is never intervened on, the corresponding column is not constrained by this argument. If each prompt has its own M_x, an anchor constrains only that prompt's column, not an unseen prompt. If the semantic map is nonlinear, finite local differences constrain sampled secants rather than one global mixing matrix. The theorem does not justify extrapolating a handful of validated edits to arbitrary natural-language variation.

A.4 A Deterministic Approximation Bound and Its Limit

To make the role of approximate selectivity explicit, retain exact semantic changes $t_k e_k$ but allow

$$\|t_k M e_k - \gamma_k e_k\|_2 \leq \epsilon_k. \tag{14}$$

Let $\mathrm{offdiag}(M) = M - \mathrm{diag}(M)$. Dropping the diagonal component in each column and summing the squared bounds gives

$$\|\,\mathrm{offdiag}(M)\|_F^2 = \sum_k \sum_{j \neq k} M_{jk}^2 \leq \sum_k \frac{\epsilon_k^2}{t_k^2}. \tag{15}$$

In particular, $|t_k| \geq t_{\min} > 0$ and $\epsilon_k \leq \epsilon$ imply a bound $\sqrt{K}\epsilon/t_{\min}$. Also $|M_{kk} - \gamma_k/t_k| \leq \epsilon_k/|t_k|$; if $s_k\gamma_k > \epsilon_k$, the diagonal entry remains positive.

This is a conditional deterministic bound, not a statistical consistency claim or an empirical certificate. If an edit instead changes $\Delta c = t_k e_k + u_k$, then

$$\|t_k M e_k - \gamma_k e_k\|_2 \leq \epsilon_k + \|M\|_2 \|u_k\|_2.$$

Off-target semantic changes can therefore be amplified by an unknown mixing matrix. Human scores of zero do not measure $\|u_k\|$ in latent units; signs do not supply $t_{\min}$. Neither the human audit nor the finite Huber loss establishes the assumptions needed to turn Eq. (15) into a numerical guarantee.

A.5 Partially Named Representations and Residual Shear

Partition semantic and nuisance variables as $(c, n) \in \mathbb{R}^K \times \mathbb{R}^J$ and let

$$\begin{bmatrix} z_s \\ z_r \end{bmatrix} = \begin{bmatrix} M_{sc} & M_{sn} \\ M_{rc} & M_{rn} \end{bmatrix} \begin{bmatrix} c \\ n \end{bmatrix} + b(x). \tag{16}$$

For a pure named edit, $\Delta n = 0$ and $\Delta c = t_k e_k$. Named-output selectivity gives $M_{sc} = D$ by the preceding proof. Requiring the residual not to move gives $M_{rc} e_k = 0$ for all k, hence $M_{rc} = 0$. No such edit tests the nuisance columns, leaving

$$M = \begin{bmatrix} D & B \\ 0 & R \end{bmatrix}, \qquad z_s = Dc + Bn + b_s(x). \tag{17}$$

Invertibility requires R to be invertible, but does not force $B = 0$. For example, $z_1 = c_1 + n$, $z_2 = c_2$, $z_r = n$ passes every pure c_1/c_2 edit while the first named coordinate still changes with nuisance. The empirical residual penalty addresses movement on observed named edits, not all possible nuisance variation. Nuisance interventions spanning the nuisance space, paired with invariant named outputs, or an independently justified structural restriction could remove B.

A.6 Finite-Feature Nullspace and Reward-Preserving Counterexample

The empirical head need not satisfy the structural relation $z = Mc + b(x)$. Let $D_h \in \mathbb{R}^{d\times m}$ contain the independent observed feature differences and partition the head as $C = [C_s^\top, c_r]^\top$. Any $U \in \mathbb{R}^{K\times d}$ with $UD_h = 0$ leaves the semantic edit predictions unchanged under $C_s' = C_s + U$. If $w_r \neq 0$, choose

$$c_r'^{\top} = c_r^\top - \frac{w_s^\top U}{w_r}. \tag{18}$$

For any h,

$$r'(h) = w_s^\top (C_s + U)h + w_r \left(c_r^\top - \frac{w_s^\top U}{w_r}\right) h \tag{19}$$

$$= w_s^\top C_s h + w_r c_r^\top h = r(h). \tag{20}$$

The residual changes on the intervention design by $-w_s^\top U D_h / w_r = 0$ as well. Thus all reward predictions and all observed intervention deltas can remain equal while unobserved semantic coordinates change.

By rank–nullity, the admissible row directions of U have dimension $d - \mathrm{rank}(D_h)$. The recorded Silver design has $d = 5120$ and rank 100, leaving 5020 such directions per semantic row. Reversing every edit adds negative columns, not new rank. These degrees of freedom establish observational underconstraint; they do not imply the trained model uses any particular shortcut. They also need not preserve $\|C\|_F$, so they are not equivalences of the weight-decayed optimization trajectory. The case $w_r = 0$ requires a different construction; Eq. (18) is not asserted there.

Relation to empirical metrics. The theorem assumes exact globally linear mixing and exact one-factor changes. Finite-sample Huber penalties and ordinal human annotations do not establish these assumptions. Coordinate scaling must be specified before combining axes into a numeric rubric. Observational natural response pairs are semantic transfer tests, not randomized interventions. We make no claim to recover a full structural causal model, uniquely identify nonlinear semantic variables, or prove that interventions are necessary for every possible identification method.

The companion source includes deterministic numerical checks of basis invariance, the named-column argument, Eq. (15), the residual counterexample, and the nullspace compensation. These checks test the algebraic implementation; they do not constitute new model-training experiments.

B Frozen-Backbone Implementation and Optimization

This section makes the parameter-efficient setting explicit and derives the quantities optimized by the released implementation. Code listings are readable adaptations of the archived source, not claims that a few lines replace the data preparation, evaluation, or provenance checks. They do not download models or launch experiments. Exact source paths and execution stages are recorded in Appendix I.

B.1 What Is Frozen and What Is Learned

The completed main real-text experiments use Qwen3-32B at the locked revision `9216db5781bf21249d130ec9da846c4624c16137`. Tokenizer output is passed through the model in evaluation and inference mode, with gradients disabled for every backbone parameter. Response-only mean pooling at block 48 gives $h_\theta \in \mathbb{R}^{5120}$. The head-training program loads cached vectors, not a differentiable backbone. Consequently no backbone activation graph or backbone optimizer state is used during head optimization. Backbone forward inference is still needed when extracting an uncached response.

Let L be the number of output coordinates and T the number of reward contexts. The bias-free projection has Ld weights; the context readouts have TL weights. Hence

$$P_{\text{head}} = Ld + TL. \tag{21}$$

The synthetic model uses $d = 64, L = 4, T = 50$, giving 456 trainable parameters over fixed generated observations. The main real-text study uses $d = 5120, L = 3, T = 1$, giving 15,363. In

the separated-ranking experiment, only two semantic rows are optimized: $2 \cdot 5120 = 10,240$. The BT reward vector and the final readout $(0, 0, 1)$ are fixed in that stage and must not be counted as newly trained parameters.

Table 4: Trainable parameter accounting for the focal models. All counts exclude frozen feature extraction and fixed normalization statistics. The synthetic control is not an LLM fine-tuning experiment.

Configuration	Projection	Readout	Total trained
Synthetic, four axes/50 contexts	$4 \cdot 64$	$50 \cdot 4$	456
Main two-axis + residual	$3 \cdot 5120$	3	**15,363**
Historical four-axis + residual	$5 \cdot 5120$	5	25,605
Separated semantic ranking	$2 \cdot 5120$	fixed	**10,240**

The main head is approximately 4.8×10^{-7} of a nominal 32-billion-parameter backbone by parameter count, or 0.000048%. This ratio uses the model's nominal scale rather than an independently counted backbone checkpoint; the head count is exact from the bias-free architecture. It is not a measured wall-clock, memory, energy, or inference-speed ratio. Full-backbone fine-tuning was not completed as a matched cost/quality baseline. The interrupted LoRA branch is a separate, incomplete experiment and is not evidence for a frozen-backbone result.

B.2 Response Boundaries, Pooling, and Cache Identity

The tokenizer applies the model's official chat template with thinking disabled. For multi-turn prompts, the preprocessing retains user/assistant roles and the latest question. Character offsets locate the response in the final assistant turn rather than a matching substring in the prompt. Only tokens fully contained in that response span enter the mask; prompt, template, and boundary-spanning tokens are excluded. With mask $m_t \in \{0, 1\}$ and block output $H_t^{(48)}$,

$$h_\theta(x, y) = \frac{\sum_t m_t H_t^{(48)}(x, y)}{\sum_t m_t}, \qquad \sum_t m_t > 0. \tag{22}$$

The model uses BF16 inference in the new-pair extractor; pooling accumulates in FP32 and stores FP16 vectors. Head training converts vectors to FP32. The maximum sequence length is 4096. If truncation is needed, the recorded policy preserves a head/tail prompt prefix within half the token budget and then the response prefix and template suffix. This policy is held fixed; the Silver cache audit reports zero truncated requests.

Cache reuse requires exact prompt/response content, response indices, model revision, token IDs, response masks, layer, and pooling identity. The main release contains 200 unique endpoint requests: 100 exact historical-cache hits and 100 newly extracted vectors. A reversed positive edit may begin with a generated response, so "original" is an orientation field, not a guarantee of a cache hit. Hashes of vectors and feature manifests are checked before head fitting. The same-pass diagnostic re-extracts both endpoints together; it does not overwrite the original TEST study.

Listing 1 shows the extraction core; source identity and token-boundary checks precede it.

B.3 TRAIN-Only Standardization and Effective Head

Let $\mathcal{T}$ be allowed TRAIN responses after human-prompt exclusions. The feature transform is

$$\mu_j = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} h_{ij}, \quad \sigma_j = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} (h_{ij} - \mu_j)^2}, \quad \bar{h}_j = \frac{h_j - \mu_j}{\max(\sigma_j, 0.1)}. \tag{23}$$

These are population-style empirical SDs (the implementation uses `ddof=0`), not trainable parameters. The same μ, σ apply to TRAIN, DEV, TEST, and new intervention endpoints. For a within-prompt pair the mean cancels, so $\Delta\bar{h} = (h_b - h_a)/\max(\sigma, 0.1)$; no separate edited-data normalization is fitted.

```python
def frozen_feature(backbone, encoded, device):
    backbone.eval().requires_grad_(False)
    ids = torch.tensor([encoded['input_ids']], device=device)
    mask = torch.tensor([encoded['response_mask']],
                        device=device, dtype=torch.bool)
    assert mask.any(1).all()
    captured = {}

    def capture(_module, _inputs, output):
        hidden = output[0] if isinstance(output, tuple) else output
        total = (hidden.float() * mask[..., None]).sum(1)
        pooled = total / mask.sum(1, keepdim=True)
        captured['feature'] = pooled.half().cpu().numpy()[0]

    handle = backbone.layers[47].register_forward_hook(capture)
    try:
        with torch.inference_mode():
            backbone(input_ids=ids, use_cache=False)
    finally:
        handle.remove()
    return captured['feature']
```

Listing 1: Frozen extraction core. The boundary-aware tokenizer supplies `encoded`; the forward hook selects block 48 (zero-based index 47).

The normalized head is $z = C\bar{h}$, $r = w^\top C\bar{h}$. In raw feature coordinates this is an affine map with slope $C\,\mathrm{diag}(\max(\sigma, 0.1))^{-1}$ and a fixed induced offset. No bias is optimized. All compared main models share the allowed TRAIN identity: 18,178 responses and 6,048 non-tie preference pairs. DEV or TEST response statistics must not enter Eq. (23).

B.4 Joint Objective and Explicit Gradients

For a preference feature difference $v_i = \bar{h}_{a_i} - \bar{h}_{b_i}$ with label b_i, write $u_i = w^\top C v_i$. The numerically stable BCE objective is

$$\mathcal{L}_{\mathrm{pref}} = \frac{1}{n_p} \sum_i \left[\mathrm{softplus}(u_i) - b_i u_i\right]. \tag{24}$$

Its per-pair gradients are

$$\nabla_C \ell_i = [\sigma(u_i) - b_i] w v_i^\top, \qquad \nabla_w \ell_i = [\sigma(u_i) - b_i] C v_i. \tag{25}$$

There is no θ update: cached v_i are constants. Although each conditional subproblem is simple, jointly optimizing C and w is bilinear and is not asserted to be convex.

For edit i, let $v_i = \bar{h}_{\mathrm{edited}} - \bar{h}_{\mathrm{original}}$, target k_i, and direction $s_i \in \{-1, 1\}$. With row $c_j^\top$ of C, the empirical terms are

$$\mathcal{L}_t = \frac{1}{n_e} \sum_i \mathrm{softplus}(-s_i c_{k_i}^\top v_i / \tau), \tag{26}$$

$$\mathcal{L}_l = \frac{1}{n_e (K-1)} \sum_i \sum_{j \neq k_i,\ j \leq K} H(c_j^\top v_i), \qquad \mathcal{L}_r = \frac{1}{n_e} \sum_i H(c_r^\top v_i), \tag{27}$$

where $H(a) = a^2/2$ for $|a| < 1$ and $H(a) = |a| - 1/2$ otherwise. This is `smooth_l1_loss` with threshold one. Its derivative is $H'(a) = \mathrm{clip}(a, -1, 1)$. For a single edit,

$$\nabla_{c_{k_i}} \ell_t = -\frac{s_i}{\tau} \sigma(-s_i c_{k_i}^\top v_i / \tau) v_i, \tag{28}$$

$$\nabla_{c_j} \ell_l = \frac{H'(c_j^\top v_i)}{K-1} v_i \quad (j \neq k_i), \qquad \nabla_{c_r} \ell_r = H'(c_r^\top v_i) v_i. \tag{29}$$

Intervention gradients update C directly, but do not update w directly; the joint preference term couples their optimization. The executed main objective is $\mathcal{L}_{\mathrm{pref}} + \mathcal{L}_t + 0.5\mathcal{L}_l + 0.25\mathcal{L}_r$, with $K = 2$

and $\tau = 1$. Dense-attribute training instead adds MSE against TRAIN-standardized correctness/coherence ratings. AdamW weight decay is an optimizer operation, not an omitted additional data-loss term.

Recorded configuration versus executed code. The V2.1 configuration contains `lambda_ortho=0.5`, inherited as metadata, but the frozen `train.py` never reads it or adds an orthogonality loss. The manuscript follows the executed code. The historical explicitly named "BT + Ortho" ablation is separate. This distinction matters when reproducing the main loss or interpreting a proposed regularization explanation.

Listing 2 gives the head and executed Silver objective.

```python
class RubricHead(nn.Module):
    def __init__(self, width=5120):
        super().__init__()
        self.encoder = nn.Linear(width, 3, bias=False)
        self.context = nn.Embedding(1, 3)
        nn.init.normal_(self.context.weight, std=0.3)

    def forward(self, difference):
        return (self.encoder(difference)
                * self.context.weight[0]).sum(-1)


def joint_loss(head, pref_dh, pref_labels, edit_dh, axes, signs):
    pref = F.binary_cross_entropy_with_logits(
        head(pref_dh), pref_labels.float())
    dz = head.encoder(edit_dh)
    i = torch.arange(len(axes), device=axes.device)
    target = F.softplus(-signs * dz[i, axes]).mean()
    leakage = F.smooth_l1_loss(
        dz[i, 1 - axes], torch.zeros_like(signs))
    residual = F.smooth_l1_loss(
        dz[:, 2], torch.zeros_like(signs))
    return pref + target + 0.5 * leakage + 0.25 * residual


# Cached, standardized features have no gradients.
# Only head parameters are optimized; no orthogonality term.
def head_optimizer(head):
    return torch.optim.AdamW(head.parameters(), lr=0.003,
                             weight_decay=0.01)
```

Listing 2: Main three-coordinate head and Silver objective. Input differences are already TRAIN-standardized. Preference-only fitting uses just BCE; dense-attribute fitting adds its separate MSE term.

B.5 Reverse Pairs and the Effective Supervision Budget

The reverse row maps (v, k, s) to $(-v, k, -s)$. Its signed target argument is unchanged: $(-s)c_k^\top(-v) = sc_k^\top v$. Since Huber loss is even, every leakage and residual term is also unchanged. With equal row weights and mean reduction, storing both directions therefore yields the same intervention objective as storing one orientation per family. It does not double the gradient weight relative to preference loss. It likewise adds no feature-difference rank and no independent observational unit. Silver-200 means 200 directed rows but 100 independent families; Silver-100 is the nested 50-family subset. This is why exclusion and uncertainty calculations operate at family/prompt level rather than directed-row level.

B.6 Independent Semantic Ranking with a Fixed Reward

The later head separates semantic fitting from preference fitting. For axis k, let $\mathcal{P}_k$ contain allowed TRAIN within-prompt pairs with unequal human rating on that axis, and let $b_{ik} = \mathbb{1}[a_{b_ik} > a_{a_ik}]$. A two-row head S minimizes

$$\mathcal{L}_{\text{rank}}(S) = \frac{1}{2}\sum_{k=1}^{2}\frac{1}{|\mathcal{P}_k|}\sum_{i\in\mathcal{P}_k}\text{BCEWithLogits}(s_k^\top v_i, b_{ik}) + \lambda\|S\|_F^2. \tag{30}$$

There are 6,060 correctness and 3,738 coherence TRAIN comparisons from 17,978 allowed responses; DEV selection uses 645 and 400 comparisons. Averaging each axis before combining them avoids letting the more numerous correctness pairs dominate solely by count. Equal-score pairs are omitted, so this objective does not impose zero off-target change. It uses more human attribute information than Silver and is not a supervision-budget-matched ablation.

For the archived BT parameters $C_{\text{BT}}, w_{\text{BT}}$, define $q_{\text{BT}} = C_{\text{BT}}^\top w_{\text{BT}}$. Attach the new semantic rows without changing reward:

$$C_{\text{sep}} = \begin{bmatrix} S \\ q_{\text{BT}}^\top \end{bmatrix}, \qquad w_{\text{sep}} = (0,0,1)^\top, \qquad r_{\text{sep}}(h) = q_{\text{BT}}^\top \bar{h}. \tag{31}$$

Only S is optimized in this stage. Reward preservation is exact algebraically and holds within 2.09×10^{-7} in saved predictions. It does not show that the semantic coordinates explain BT preference, since their reward coefficients are zero. The advantage supported here is a small semantic interface that can be improved while leaving the existing reward intact.

Listing 3 separates semantic optimization from the fixed reward attachment.

```python
def ranking_loss(semantic, dh, human_delta, l2=0.1):
    logits = semantic(dh)  # Linear(5120, 2, bias=False).
    losses = []
    for k in range(2):
        keep = human_delta[:, k] != 0
        assert keep.any()
        label = (human_delta[keep, k] > 0).float()
        losses.append(F.binary_cross_entropy_with_logits(
            logits[keep, k], label))
    return (sum(losses) / 2
            + l2 * semantic.weight.square().sum())


def attach_fixed_reward(semantic_weight, bt_encoder, bt_readout):
    q_bt = bt_encoder.T @ bt_readout
    encoder = torch.cat([semantic_weight, q_bt[None, :]], dim=0)
    readout = encoder.new_tensor([0.0, 0.0, 1.0])
    return encoder, readout
```

Listing 3: Separated ranking objective and attachment to a fixed BT reward. Regularization is selected by DEV ranking loss over the recorded grid; this code does not perform that selection.

The recorded grid $\lambda \in \{0, 0.1, 1\}$ produces 15 fits (three values, five seeds). Mean DEV ranking loss selects 0.1, yielding $72.67 \pm 0.11\%$ DEV composition. The $+3.81$-point comparison to HumanDelta-800 is a post-selection development result with extra labels; the $+0.73$-point comparison to aligned BT is inconclusive. No original-TEST confirmation is available for this selected head. All these qualifications are necessary when describing the parameter-efficient gain.

C Evaluation Definitions and Uncertainty Calculations

This section specifies the completed main two-axis evaluation. The historical four-axis and synthetic protocols differ in normalization, alignment, and ranking population; their conventions remain in Appendix F. They are not silently converted into the main protocol. All rules below are read from the archived evaluator rather than inferred from rounded summary tables.

C.1 Freezing Mappings and Scales Before TEST

For an unnamed BT head, let R^{DEV}_{jk} be the Spearman correlation between its first two coordinates and the two human attributes on DEV. The mapping maximizes total absolute correlation,

$$\pi^* = \arg\max_{\pi \in S_2} \sum_{k=1}^{2} |R^{\mathrm{DEV}}_{\pi(k),k}|, \qquad \eta_k = \mathrm{sign}(R^{\mathrm{DEV}}_{\pi^*(k),k}). \tag{32}$$

The implementation uses Hungarian assignment and takes a nonnegative correlation to have positive sign. Named models keep their requested order and signs. Undefined DEV alignment is an error, not a reason to choose a mapping using TEST. The aligned coordinate is $\hat{z}_k = \eta_k z_{\pi^*(k)}$.

Model and human attribute scales are then fitted on allowed TRAIN responses:

$$\tilde{z}_k = \frac{\hat{z}_k - \mu_{z,k}}{\max(\sigma_{z,k}, 10^{-8})}, \qquad \tilde{a}_k = \frac{a_k - \mu_{a,k}}{\max(\sigma_{a,k}, 10^{-8})}. \tag{33}$$

The 10^{-8} denominator floor here is distinct from the feature-standardization floor 0.1 in Eq. (23). All mappings and scale files are saved and hashed before computing TEST outcome metrics. Raw BT is retained only as an additional diagnostic view of the same five fits. Comparing a named model only to arbitrary raw signs would confound semantic learning with alignment.

Positive diagonal indeterminacy has an operational consequence: a weighted rubric needs a scale convention. TRAIN normalization removes a fixed positive scale in ideal nondegenerate coordinates, but does not remove nuisance contamination, prompt-specific transformations, or off-target mixing. Scale standardization is therefore part of the scoring protocol, not a proof of identification.

C.2 Preference, Purity, Natural Effects, and Stability

For each non-tie preference label b_i, define logit $u_i = r(x_i, y_{a_i}) - r(x_i, y_{b_i})$. Preference accuracy is

$$\mathrm{Acc}_{\mathrm{pref}} = \frac{1}{n_p} \sum_i \mathbb{1}[u_i(2b_i - 1) > 0]. \tag{34}$$

Model ties are incorrect, not half-correct. The main TEST denominator is 352 non-tie preference pairs; it is not all 1,038 responses or all within-prompt pairs.

On TEST responses, let $R_{jk} = \mathrm{Spearman}(\tilde{z}_j, a_k)$. The full 2×2 matrix is retained, and the two-axis purity margin is

$$\mathrm{Purity} = \frac{1}{2}\left(|R_{11}| - |R_{12}| + |R_{22}| - |R_{21}|\right). \tag{35}$$

The margin can be negative. High individual correlations or stable axes do not imply distinct criteria: both axes may track similar holistic quality. Taking absolute correlations also does not certify the intended direction; signed intervention effects provide a separate diagnostic. Undefined correlations are missing values, not zeros.

A strict natural pair belongs to criterion k if it shares one prompt, satisfies $|a_{b,k} - a_{a,k}| \geq 1$, and has zero differences on all three other rated attributes. For orientation $s_i = \mathrm{sign}(a_{b,k} - a_{a,k})$, define

$$e_i = s_i(\tilde{z}_{b,k} - \tilde{z}_{a,k}), \qquad \mathrm{TE} = \frac{1}{n_{\mathrm{nat}}} \sum_i e_i. \tag{36}$$

The main strict set has 94 correctness and 17 coherence pairs, so TE is pair-weighted across 111 pairs. It is not an equally weighted mean of criterion-specific effects. The relaxed set keeps only the other named criterion unchanged and has 195/37 pairs (232 total). Both supports are reported. Natural pairs isolate changes in recorded ratings but are observational, not randomized interventions or exact latent one-factor changes.

For five fitted seeds, stability averages the Pearson correlation of each corresponding TEST coordinate over the ten unordered seed pairs:

$$\mathrm{Stability} = \frac{1}{2\binom{5}{2}} \sum_{k=1}^{2} \sum_{s<t} \mathrm{Corr}(\tilde{z}^{(s)}_k, \tilde{z}^{(t)}_k). \tag{37}$$

This is one joint statistic. The ten comparisons share fitted models and are not ten independent training replicates. The main stability entries consequently have no invented per-fit SD.

C.3 Rubric Composition: Order of Aggregation Matters

Use the fixed set $\mathcal{W} = \{(t, 1-t) : t = 0, 0.05, \ldots, 1\}$ with $|\mathcal{W}| = 21$. For rubric $v \in \mathcal{W}$, response scores are $s_z = v^\top \tilde{z}$ and $s_a = v^\top \tilde{a}$. Enumerate all unordered response pairs within each TEST prompt. For pair i define

$$V_{iv} = \mathbb{1}[|\Delta s_{a,iv}| > 10^{-8}], \qquad B_{iv} = \mathbb{1}[\Delta s_{z,iv} \Delta s_{a,iv} > 0]. \tag{38}$$

Composition is computed as

$$A_v = \frac{\sum_i V_{iv} B_{iv}}{\sum_i V_{iv}}, \qquad \text{Composition} = \frac{1}{21} \sum_{v \in \mathcal{W}} A_v. \tag{39}$$

Thus valid pairs are pooled within a rubric, and rubrics receive equal weight afterward. It is not an equal-prompt average, nor one pooled accuracy across all rubric–pair combinations. Prompts with more response pairs contribute more within a rubric. Human ties are omitted separately for each rubric; model ties count as incorrect. Denominators are stored for every rubric and seed.

Within-prompt Kendall τ_b and Spearman ρ are computed for each rubric and then averaged over defined prompts, followed by the rubric average. Constant/tied groups with undefined statistics are omitted with their support recorded. For a two-response prompt, either coefficient reduces to sign agreement when neither score is tied. The historical protocol's population-wide rank correlations measure something different and are not used as matched comparisons here.

C.4 Paired Prompt-Cluster Bootstrap

Let $p = 1, \ldots, P$ index all 518 TEST prompts. For seed s and rubric v, store the prompt-level correct and valid pair counts $N_{spv}^{(m)}$ and D_{spv} for method m. Draw

$$(c_1^{(b)}, \ldots, c_P^{(b)}) \sim \text{Multinomial}(P; 1/P, \ldots, 1/P) \tag{40}$$

for $b = 1, \ldots, 10{,}000$, using seed 73019. The same prompt multiplicities are used for both methods and every one of the five fixed seed fits. For composition,

$$T_m^{(b)} = \frac{1}{5} \sum_{s=1}^{5} \frac{1}{21} \sum_v \frac{\sum_p c_p^{(b)} N_{spv}^{(m)}}{\sum_p c_p^{(b)} D_{spv}}, \qquad \delta^{(b)} = T_{\text{Silver}}^{(b)} - T_{\text{BT}}^{(b)}. \tag{41}$$

The percentile interval is the 2.5th and 97.5th percentiles of finite $\delta^{(b)}$ values. Preference and TE use analogous resampled sums divided by their own resampled supports. Spearman-based purity requires recomputing ranks with repeated prompt observations; resampling already-computed per-prompt correlations is not equivalent. The evaluator uses weighted midranks, including ties, to reproduce that operation.

Listing 4 implements the composition aggregation using the same multiplicities for both methods.

Reported seed SDs instead use

$$\text{SD}_{\text{seed}}(T) = \sqrt{\frac{1}{4} \sum_{s=1}^{5} (T_s - \bar{T})^2}.$$

They describe variation across the five fits and are not standard errors or confidence intervals. The bootstrap conditions on those fits, the TRAIN scales, and DEV mappings; it does not resample seeds, refit models, or cover repeated DEV model selection. Intervals are unadjusted across endpoints. The main Silver composition interval is entirely negative, whereas the separated-ranking versus BT DEV interval includes zero; neither conclusion should be replaced by a favorable point estimate alone.

C.5 Blind Orientation and Human-Audit Uncertainty

For each blind pair, let $o = +1$ when B is edited and A is original, and $o = -1$ otherwise. If the annotator reports d_j^{B-A}, restore

$$\Delta a_j^{\text{edited}-\text{original}} = o \, d_j^{B-A}. \tag{42}$$

```python
import numpy as np


def composition_draws(numerator, denominator, counts):
    correct = np.einsum('bp,spr->bsr', counts, numerator)
    support = np.einsum('bp,spr->bsr', counts, denominator)
    rate = np.divide(correct, support,
                     out=np.full_like(correct, np.nan),
                     where=support > 0)
    return np.nanmean(rate, axis=2).mean(axis=1)


def paired_interval(base_num, target_num, denominator,
                    draws=10000, seed=73019):
    assert base_num.shape == target_num.shape == denominator.shape
    prompts = base_num.shape[1]
    rng = np.random.default_rng(seed)
    counts = rng.multinomial(prompts, np.full(prompts, 1 / prompts),
                             size=draws)
    # Share prompt counts across methods and fixed seeds.
    base = composition_draws(base_num, denominator, counts)
    target = composition_draws(target_num, denominator, counts)
    effect = target - base
    return np.nanquantile(effect, [0.025, 0.975]), effect
```

Listing 4: Paired composition bootstrap over prompt aggregates. Five seed fits remain fixed; both methods share the same resampling counts. The full evaluator also handles other endpoints and weighted rank recomputation.

The historical pass is $s\Delta a_k \geq 1$ and $\Delta a_j = 0$ for all $j \neq k$. The V2 pass adds `other_meaningful_change=no`. Applying the appropriate rule after orientation restoration yields the frozen audit counts. Confidence is not a replacement label, and free-text explanations cannot be direction-corrected by negating a string. Their orientation must remain explicit or be reviewed and rewritten separately.

For x passes in n reviewed pairs, the plotted Wilson interval uses $z = 1.96$, $\hat{p} = x/n$, and

$$\frac{\hat{p} + z^2/(2n) \pm z\sqrt{\hat{p}(1-\hat{p})/n + z^2/(4n^2)}}{1 + z^2/n}. \tag{43}$$

The three V2 denominators are 20 overall, 16 high-ranked, and four boundary examples. The overall set contains the other two groups; they are not three independent experiments. The design deliberately selects a boundary example per cell, so these intervals do not correct selection bias or estimate the precision of an arbitrary deployment pool. The 16/16 result has lower bound 80.6%, not proof of perfect precision. The subsequent V2.1 guard requires its own independent human audit.

D Human Audit Protocol and Scope

The original 80-pair audit is balanced over four criteria and two requested directions, ten per cell. Each prompt and reverse-edit family appears once. Sampling uses TRAIN candidates that passed structural screening, enriched with 56 borderline examples (including single-judge-clean examples without a second independent judge) and 24 likely-impure examples. No dual-judge high-confidence subset was available. These design choices preclude reading Table 7 as the precision of an independently certified full training pool.

The annotator sees prompt, response A, response B, and blank rating fields. The requested criterion, direction, generator, scores, and original/edited identity are hidden. A/B order is randomized. Ratings always describe B relative to A; an orientation key converts the four numeric differences back to original-to-edited before scoring. Raw free-text reasons remain explicitly labelled in their original A/B orientation rather than being naively sign-flipped. The four dimensions use the following operational definitions:

- **Correctness:** factual, logical, mathematical, or task-specific correctness; stylistic improvements alone do not qualify.
- **Coherence:** discourse flow, reference continuity, and internal consistency; an isolated factual error need not change discourse coherence.
- **Complexity:** conceptual or reasoning sophistication, not merely formal vocabulary or length.
- **Verbosity:** redundant or nonessential elaboration, not useful added reasoning.

Each signed score lies in $\{-2, -1, 0, 1, 2\}$. The historical four-attribute pass requires the target to move at least one point in the requested direction and every other rated dimension to be zero. This frozen definition does not use the other-meaningful-change flag. One coherence-negative example passes the four-dimensional test but has that flag set; adding the stricter V2 requirement changes coherence from 18/20 to 17/20 (negative: 9/10 to 8/10), leaving the other criteria unchanged. That example is also in the historical TinyHuman calibration set, so that training set is not described as V2-certified. Confidence and short reasons support auditability but do not create independent raters. The current artifacts do not establish inter-annotator agreement.

The sign-normalized mean change matrix is

	C	Coh	$Comp$	$Verb$
$do(C)$	1.15	0.10	0	0
$do(Coh)$	0.15	1.75	0	0
$do(Comp)$	0	−0.05	0.60	0.20
$do(Verb)$	0	−0.90	0	1.05

where each row averages $s\Delta a$ over 20 annotated pairs. This is a descriptive matrix of judged changes, not an identified causal mixing matrix or a proof of intrinsic coupling between concepts.

D.1 Direction-Specific Support

Criterion	Negative passes	Positive passes
Correctness	8/10	9/10
Coherence	9/10	9/10
Complexity	0/10	3/10
Verbosity	0/10	2/10

Both directions have examples passing the historical four-attribute criterion in this sample. This contrasts with the historical accepted pool, whose main training set contains no coherence-positive anchor, but does not establish a matched comparison between generators.

The earlier two-axis follow-up trains the TinyHuman diagnostic on 17 calibration examples passing the historical criterion (correctness negative/positive: 4/4; coherence negative/positive: 4/5). Its human evaluation contains 18 locked pairs meeting that same criterion. Natural support is 111 strict four-dimension pairs; relaxing the requirement to the other named dimension yields 232 pairs and is secondary. No locked prompt enters TinyHuman training.

D.2 Fresh V2 Blind Test

The new 20 examples have zero candidate overlap with prior human annotations and use 20 distinct prompts and families. Each correctness/coherence direction cell contributes four high-ranked and one boundary example. Strict counts are 4/5 for correctness-positive and 5/5 for each other cell. The high subset passes 16/16; the boundary subset passes 3/4. This is one quota-stratified human validation set, not 20 repeated validation runs.

The correctness-positive failure concerns a factual correction within an exercise that permits false statements. For example, replacing a deliberately false statement with a true one need not improve compliance with the exercise. This motivates the V2.1 task-relative rule, but the original label and denominator remain unchanged. Neither this failure analysis nor the rule change licenses relabelling the V2 observation as a V2.1 validation.

E Filter Versions and Silver Construction

Keep the versions separate. The archived first two-axis filter (P0/R1) accepts seven of its 20 locked examples and has six successes under the historical four-attribute criterion; it fails coverage and the main training branch does not proceed. V2 denotes the subsequent direct GPT-6 review tested on a fresh 20-pair blind packet, with the additional no-other-meaningful-change requirement. V2.1 is the later task-relative revision. These version labels refer to different stages, not repeated measurements of a single unchanged filter.

Task-relative guard. The V2.1 protocol states:

> Correctness is defined relative to the task objective and prompt constraints, not factual truth in isolation. If the task explicitly permits false statements or asks to transform/identify false content, factual truth alone does not determine correctness.

The rule is applied to new candidates only. It is not used to retrospectively change V2's observed 19/20, 16/16, or 3/4 counts.

Selection and independence unit. The current pool contains 800 directed rows, 400 reverse-pair families, and 200 rows per criterion–direction cell. Forty rows overlap human-reviewed families/prompts and are excluded before screening. The remaining 380 families receive direct offline model review. Ordinal scores prioritize clear selective changes: 95 and 85 are eligible; 75 is boundary and excluded from the main set; lower scores indicate unresolved, off-target, or invalid changes. Ranking uses score followed by a frozen hash tie-break. Final selection enforces distinct prompts across families.

The high-screen pool has 74 correctness and 117 coherence families, insufficient for a balanced 100-per-cell second tier without relaxing eligibility. The main selection uses 50 families per criterion, each with both byte-verified directions: 200 rows from 100 independent pairs. Of the 200 rows, 186 have score 95 and 14 score 85. All selected families undergo full-context inspection; one prospective family is downgraded and replaced. The release includes nested Silver100, with 25 rows in each direction cell and 50 independent families. Do not split reverse rows across partitions or count them as independent bootstrap observations.

Artifact and leakage checks. Each training JSONL record contains exactly `prompt`, `original_response`, `edited_response`, `criterion`, `direction`, `candidate_id`, and `gpt6_score`. The score is not a target; explanations are excluded. A separate manifest records provenance, family/prompt identities, and text hashes. Human candidate, family, and prompt overlap is zero for both the old 80 and new 20. Original held-out prompt exclusions are retained. Source-table SHA256 values match the archived split provenance; this is not a new independent decoding audit of the parquet splits.

Completed matched comparison. The fixed follow-up compares Latent-BT-2, Clean-ILR-2, Full-Attribute-2, and the nested Silver100 ablation over seeds 0–4, frozen Qwen3-32B block-48 response-mean features, a linear three-coordinate head, and common preference splits. Full-Attribute has extra dense human labels. All historical-human prompts are removed from preference and attribute training and normalization consistently across models. The latest snapshot records 20 completed fits; the earlier data-release integrity file records preparation-time status only.

Features are keyed by exact prompt/response content: reversing an edit makes a synthetic response the *original* side of the positive pair. The 200 training rows imply 200 unique prompt–response feature requests. The completed cache audit verifies 100 exact historical-cache matches and 100 newly extracted vectors, with no truncated intervention requests. A later diagnostic re-extracts both endpoints together and is kept separate from the frozen main TEST evaluation.

Completed endpoints are preference accuracy, human axis purity, natural target effect (strict four-dimensional support primary), two-dimensional composition over 21 fixed weights, and five-seed stability. Alignment and scaling are fixed on TRAIN/DEV. The 10,000-replicate paired bootstrap

resamples prompt clusters and averages the five fixed fitted seeds, without treating reverse directions as independent. Appendix H gives the completed results and distinguishes later DEV exploration.

F Additional Completed Results and Metric Conventions

Synthetic setup and controls. The data seed is 20260915. Factors are standard Gaussian, the observation mixing matrix has orthogonal columns of norm $\sqrt{64}$, and preference inverse temperature is two. The 50 contexts combine dense, sparse signed, and Dirichlet weights. Seeds 0–4 share the dataset but vary initialization and nested intervention samples. Synthetic training uses 800 full-batch AdamW steps, learning rate .02, weight decay .0001, $\lambda_t = 1$, $\lambda_l = .5$, and no residual penalty. True probe labels do not select an ILR checkpoint. Completed work includes eight budgets, off-target-noise checks, orthogonality fits, and single-seed shuffled-label/sign-flipped controls. Those controls are not five-seed comparisons.

Synthetic axis recovery is the mean absolute Pearson correlation with the generating factors. The preference-only baseline receives optimal Hungarian matching on TEST probes for this descriptive recovery metric; intervention models retain their named coordinate order. Subspace recovery is the mean canonical correlation between centered learned and true probe representations. These differ from the real-text Spearman purity margin. Synthetic composition uses DEV-fixed baseline matching/signs, TEST standardization, 64 fixed rubric weights, and 10,000 fixed random probe pairs; its normalization is likewise conditional on the held-out probe population.

Historical real setup. The main four-criterion model uses $\lambda_t = 1$, $\lambda_l = .5$, $\lambda_r = .25$, $\tau = 1$, learning rate .003, and weight decay .01. The tested optimization protocol permits 600 steps and uses DEV preference-based checkpoint selection; the implementation/configuration archive records exact per-run stopping. The main budget $N = 50$ is per criterion, not per direction. Accepted TRAIN support is 159 correctness, 71 coherence, 93 complexity, and 253 verbosity edits; their negative/positive counts are 150/9, 71/0, 19/74, and 55/198. The natural human-pair diagnostic has 156 pairs (94/17/13/32 by criterion), allowing non-target differences up to .5, unlike the strict equality used in the two-axis follow-up.

Historical method	Preference acc.	Axis purity	Composition acc.
Scalar-BT	$.7409 \pm .0088$	–	–
Latent-BT	$.7443 \pm .0088$	$-.0661 \pm .0223$	$.6061 \pm .0068$
Latent-BT + Ortho	$.7398 \pm .0091$	$-.0131 \pm .0130$	$.6000 \pm .0098$
4D ILR	$.7318 \pm .0308$	$-.0101 \pm .0213$	$.5667 \pm .0437$
Dense Attribute+BT	$.7034 \pm .0232$	$.1468 \pm .0113$	$.6510 \pm .0083$

All $\pm$ entries here are sample SD across five fits. Historical unlabelled-baseline recovery and purity use optimistic TEST Hungarian matching and are descriptive diagnostics. Composition and intervention effects instead use DEV-fixed matching and signs. No claim that all columns share the same alignment procedure is intended.

Stability and ranking caveats. The historical main stability statistic is .8341 for Latent-BT and .5037 for ILR, but compares five absolutely matched baseline axes with four signed named ILR axes. An equal-four-coordinate, DEV-fixed analysis yields .8852 versus .5037; high baseline correlation may reflect repeated holistic-reward directions rather than distinct semantic variables. Seed-partner statistics share models and are not independent replicates. Historical composition Kendall/Spearman rank across the TEST population, whereas the two-axis diagnostic ranks within prompt; their coefficients must not be treated as a common benchmark.

Uncertainty. The historical ILR-minus-Latent-BT composition difference is $-.039485$, with 10,000 paired prompt/seed bootstrap replicates and 95% interval $[-.080804, -.002757]$. Preference difference is $-.0125$, with interval $[-.047731, .022663]$. The axis-purity difference .055991 has a conditional interval $[-.004513, .081763]$ from 1,000 prompt-cluster replicates, so a reliable purity gain is not established. These intervals are unadjusted across outcomes, conditional on saved fits, fixed rubrics, and the recorded normalization. They do not cover new rubric populations, all data-collection uncertainty, or full-pipeline retraining.

TinyHuman is exploratory. Five-seed TinyHuman-ILR-2 results are: preference $.7472 \pm .0045$; locked-human TE $1.0727 \pm .2083$; strict natural TE $.0908 \pm .1304$; purity $-.0184 \pm .0299$; composition accuracy $.5997 \pm .0977$; within-prompt Kendall $.1982 \pm .1989$; Spearman $.1985 \pm .1990$. Stability .7360 is a joint five-seed statistic rather than a mean of independent per-seed estimates. No corresponding formal two-axis baseline or dense-reference run is available in the archived branch.

G Numerical Results and Human-Validation Counts

The main text reserves quantitative figures for controlled identification and human validation (Figures 3–4). The following tables retain exact values and original definitions, including the historical negative result summarized in the main text. Synthetic and historical-model errors are sample SD over five fitted seeds; human-validation intervals are Wilson 95% confidence intervals, not seed variability.

Table 5: Synthetic positive control, mean $\pm$ sample SD over five seeds. N counts interventions per criterion. Exact anchors improve axis recovery and standardized composition, not merely preference fit.

N	Preference accuracy	Axis recovery	Composition accuracy
0	$.8219 \pm .0008$	$.7300 \pm .0724$	$.7718 \pm .0447$
1	$.8246 \pm .0007$	$\mathbf{.9963 \pm .0013}$	$\mathbf{.9755 \pm .0038}$
5	$.8246 \pm .0004$	$.9995 \pm .0000$	$.9909 \pm .0007$
25	$.8244 \pm .0006$	$\mathbf{.9998 \pm .0000}$	$\mathbf{.9939 \pm .0002}$

Table 6: Historical four-axis experiment, five-seed mean $\pm$ sample SD. TE is normalized by TRAIN latent SD. The two TE columns evaluate different pair distributions; composition evaluates human ratings.

Method	Preference acc.	Generated TE	Natural TE	Composition acc.
LATENT-BT	$.7443 \pm .0088$	$.0816 \pm .0096$	$.2408 \pm .0375$	$.6061 \pm .0068$
4D ILR	$.7318 \pm .0308$	$1.1129 \pm .0898$	$.1220 \pm .1327$	$.5667 \pm .0437$

Table 7: Human80 diagnostic audit. The historical one-factor criterion requires a target hit and zeros on the other three rated dimensions; it does not use the additional-change flag. Counts describe the sampled construction, not intrinsic intervenability.

Criterion	Target hit	Four-attribute pass	Pass rate
Correctness	19/20	**17/20**	**85%**
Coherence	20/20	**18/20**	**90%**
Complexity	12/20	3/20	15%
Verbosity	20/20	2/20	10%

Table 8: Fresh V2 blind validation. The 16 high-ranked and four boundary pairs were sampled by quota. Intervals describe uncertainty within these strata; they do not eliminate selection bias or validate the later V2.1 rule.

Stratum	Strict passes	Observed rate	Wilson 95% CI
All reviewed	**19/20**	**95.0%**	76.4–99.1%
High-ranked	**16/16**	**100.0%**	80.6–100.0%
Boundary	3/4	75.0%	30.1–95.4%

H Completed Two-Axis Study and Post-Study Diagnostics

This appendix reports the update archived at `clean2d_updates/20260925T182503Z`. Experiment labels V2.1–V2.5 identify successive branches of the archive; they do not denote repeated

independent confirmations. The V2.1 TEST evaluation is the completed main comparison. Subsequent branches investigate that result on DEV and do not reevaluate the original TEST.

H.1 Main Comparison: Protocol and Exact Results

All four configurations use frozen Qwen3-32B block-48 response-mean features, a three-coordinate linear head, learning rate 0.003, weight decay 0.01, and the same DEV preference-loss checkpoint selection. The budget is at most 600 steps, evaluated every ten steps, with patience 100 and minimum 100 steps. The target, leakage, and residual coefficients are 1, 0.5, and 0.25. Although the configuration records `lambda_ortho=0.5`, the frozen V2.1 training source does not add an orthogonality term; it is not part of the executed objective. Feature standardization uses allowed TRAIN statistics with SD floor 0.1. All methods share 18,178 allowed TRAIN responses and 6,048 non-tie preference training pairs after excluding the 100 historical-human prompts. Scores and review explanations are not model inputs. Dense attribute training receives additional TRAIN correctness/coherence labels; it is neither supervision-budget matched nor a guaranteed upper bound.

BT's first two coordinates are permuted and signed using DEV human-score correlations, then frozen for TEST. Named models are not rematched. Human and latent scales for composition use allowed TRAIN only. For each of 21 weights $(t, 1-t)$, $t \in \{0, .05, \ldots, 1\}$, composition computes within-prompt pair accuracy and then averages across rubrics. Human differences at most 10^{-8} are omitted; model ties count as errors. Kendall τ_b and Spearman ρ are averaged within prompt and rubric, omitting undefined tied/constant groups with recorded support. The original TEST has 518 prompts, 1,038 responses, and 352 non-tie preference pairs. Natural strict support is 111 pairs (94 correctness, 17 coherence); relaxed support is 232 (195 and 37). Target effect is pair-weighted, not an equal-weight average of the two criteria.

Table 9: Completed original-TEST comparison. Means $\pm$ sample SD over five fits; preference and composition are percentages. Stability is one joint statistic across ten seed pairs, not an independent-fit mean with a seed SD. BT denotes DEV-aligned coordinates.

Model	Preference	Purity	Natural TE	Composition	Stability
BT-2	$74.26 \pm .65$	$-.003 \pm .005$	$.272 \pm .026$	$72.57 \pm .76$	.906
Silver-200	74.03 ± 3.17	$-.009 \pm .009$	$.107 \pm .106$	61.40 ± 8.58	.807
Silver-100	73.98 ± 3.02	$-.022 \pm .018$	$.123 \pm .092$	62.78 ± 6.45	.850
Full-Attribute	$74.49 \pm .74$	$.048 \pm .018$	$.181 \pm .044$	66.63 ± 1.34	.567

The unmapped BT diagnostic has composition $.5451 \pm .2469$, natural TE $.0543 \pm .2900$, and stability $-.1904$. It uses the same five fits, not five additional runs. Comparing only to this arbitrary-sign view would overstate intervention benefits. Conversely, high composition after DEV alignment does not itself establish distinct semantic axes: aligned BT's purity remains near zero.

Table 10: Paired Silver-200 minus aligned BT differences on the original TEST. All values are on the unit scale, not percentage points. Intervals use 10,000 paired prompt-cluster resamples conditional on five fitted seeds, TRAIN scales, and DEV mappings.

Metric	Difference	95% percentile interval
Preference accuracy	$-.002273$	$[-.022961, .018824]$
Axis purity	$-.006522$	$[-.047279, .033247]$
Strict natural TE	$-.165437$	$[-.313099, -.028125]$
Relaxed natural TE	$-.217574$	$[-.341687, -.093573]$
Composition accuracy	$-.111722$	$[-.148992, -.074427]$
Within-prompt Kendall	$-.222520$	$[-.297930, -.147623]$
Within-prompt Spearman	$-.222745$	$[-.298015, -.147772]$

The same resampled prompts are used for every model and all five fixed fits; seeds are not resampled. Rank-based purity recomputes Spearman midranks on repeated sampled observations. Intervals are not corrected for multiple outcomes and do not cover retraining or later DEV model selection. Silver-200 composition is lower than its paired BT fit for each of the five seeds. The nested Silver-100 comparison is descriptive; no corresponding paired significance claim is made. Relaxed TE

means are .4473 for BT, .2297 for Silver-200, .2467 for Silver-100, and .3207 for Full-Attribute. Corresponding within-prompt Kendall/Spearman means are .4535/.4538, .2310/.2311, .2577/.2579, and .3348/.3351.

H.2 Why Fitting the Edit Labels Is Insufficient

Independent CPU replay reproduces selected-checkpoint DEV preference predictions within 2×10^{-7}. Both Silver criteria have 100% training direction accuracy. Extending optimization to step 599 changes mean DEV composition from 59.74% to 60.60%, while preference loss rises from .5287 to 2.9419; even post-hoc selection by each seed's best DEV composition yields only 61.03%. Positive semantic-readout initialization increases composition to 64.40%, largely through seeds whose initial signs were negative, without improving purity. These are post-study diagnostics, not alternative selections for the original TEST table.

Same-pipeline extraction removes the historical batching difference between cached and edited endpoints. Retraining gives $59.72 \pm 8.25\%$ DEV composition versus $59.74 \pm 8.16\%$ with original endpoints. When 20 edit families are excluded from intervention training, the comparison is $58.38 \pm 9.05\%$ versus $58.36 \pm 9.04\%$, with perfect held-out direction accuracy. Only this latter branch is an intervention holdout; its prompts still occur in preference TRAIN. The all-family branch's similarly named subgroup is in-sample. These results narrow the feature-provenance explanation but do not constitute a new TEST evaluation.

The correctness construction admits $f(y) = -\sum_{v \in \text{integers}(y)} v$, which predicts all 100 directed correctness labels; coherence edits leave this value unchanged. This is an existence demonstration of a template shortcut, not model-attribution evidence. A second diagnostic writes the three-coordinate head as semantic rows C_s and residual row $c_r^\top$. Let D contain the 100 independent standardized feature differences, so $\text{rank}(D) = 100$ in dimension 5,120. Any perturbation U satisfying $UD = 0$ preserves semantic intervention deltas. For nonzero residual readout w_r, setting

$$C_s' = C_s + U, \qquad c_r'^{\top} = c_r^\top - \frac{w_s^\top U}{w_r} \tag{44}$$

also preserves the full reward function and residual intervention deltas. Saved constructions change DEV composition while retaining all checked rewards and intervention-coordinate differences within 10^{-8}. Their parameter norms differ, so regularization need not treat them as equivalent. This empirical underconstraint is broader than the theorem's restricted global semantic mixing model.

H.3 Natural Human Contrasts and Separated Semantic Heads

The natural-contrast branch uses target human-score difference at least one and zero differences on the other three rated attributes. Equal observed ratings do not guarantee equal latent properties; these are not randomized one-factor interventions. The 200/800 directed training rows correspond to 100/400 independent families and distinct prompts. A further 100 families (50 per axis) are excluded from preference, attribute, intervention, and normalization training for every method. All five configurations use positive initialization and DEV preference-loss selection, giving 25 completed fits.

Table 11: Post-study exploration, not original-TEST confirmation. DEV percentages are five-fit means ± SD. Held-out direction accuracies use 50 excluded natural families per criterion. Extra human attribute supervision differs across methods.

Method	DEV preference	DEV composition	Held-out C	Held-out Coh
Aligned BT	$75.14 \pm .53$	$71.94 \pm .56$	61.2%	49.6%
Silver-200	74.77 ± 1.32	64.45 ± 1.90	61.2%	58.8%
HumanDelta-200	$74.89 \pm .42$	66.85 ± 2.01	57.2%	53.6%
HumanDelta-800	$75.05 \pm .90$	68.86 ± 1.34	57.6%	54.4%
Full-Attribute	$76.21 \pm .91$	65.60 ± 1.90	61.2%	53.6%

HumanDelta-800 minus BT has held-out correctness/coherence direction differences $-3.6/+4.8$ points with 95% intervals $[-15.2, 7.2]/[-8.0, 17.6]$. Against Silver the differences are $-3.6/-4.4$

with intervals $[-15.6, 8.4]/[-20.8, 11.6]$. Each interval uses 10,000 paired prompt resamples within its criterion, conditional on five fits. No consistent gain is established by these unadjusted intervals.

The separated-head branch fits semantic coordinates without preference gradients. Dense and contrast ridge heads reach 69.70% and 65.65% DEV composition (alpha 10,000, selected by DEV MSE); these are deterministic semantic fits attached to five BT reward heads, not five independently trained semantic replicas. Ranking heads use all 17,978 allowed TRAIN responses, with 6,060 correctness and 3,738 coherence non-tie within-prompt comparisons. Equal-score pairs provide no ranking label and do not enforce zero coordinate change. Three L2 values $\{0, .1, 1\}$ are fitted over five seeds; mean DEV ranking loss selects .1, giving $72.67 \pm .11\%$ composition and purity .0190. Preference is preserved by the construction

$$z(h) = [z_C(h), z_{Coh}(h), q_{\mathrm{BT}}^{\top} h]^{\top}, \qquad w = (0, 0, 1)^{\top}, \tag{45}$$

with saved reward discrepancies at most 2.09×10^{-7}. Thus the semantic heads do not explain reward. Paired DEV intervals give ranking minus BT $+.73$ points $[-1.27, 2.72]$, and ranking minus HumanDelta-800 $+3.81$ points $[1.51, 6.07]$. These are conditional diagnostics after model selection, not independent confirmatory intervals.

H.4 Broader Search and Incomplete Adaptation

The next screen contains 32 tree settings, 48 linear settings, eight MLP settings, six structured settings, and six corrected ensembles (100 total), plus three cross-prompt settings. This is seed-0 exploration, not five-seed confirmation. The best DEV score is 72.7405%, the same block-48 mean-pooling linear configuration as the selected ranking model's seed 0. Best MLP and structured scores are 72.7068% and 72.6403%; the corrected ensemble reaches 72.2021%. The old ensemble artifact omitted final normalization by the ensemble's own TRAIN SD and is superseded; only corrected scores are used here. Many nonlinear heads warm-start the selected linear head, and structured training directly uses the same 21 rubric weights, precluding an unseen-weight transfer claim.

The three cross-prompt configurations add 99,990 sampled TRAIN comparisons and reach 71.7861%, 72.5729%, and 72.5860%. They additionally assume comparability of ratings across prompts. None of these 103 settings reaches the exploratory 80% target. Repeated DEV use and extra supervision limit what can be inferred from the best value.

Finally, four LoRA configurations were interrupted and their processes terminated before completion. Archived curves contain steps 0 and 100 of a planned 400, but no final metrics or completed TRAIN-normalized composition. Intermediate checkpoints do not support a final accuracy claim. This branch is recorded for completeness and excluded from result tables; it does not establish whether representation adaptation succeeds or fails.

I Experimental Workflow, Execution Ledger, and Reproduction Guide

The work comprises data construction, blinded calibration, immutable feature preparation, controlled fitting, independent scoring, and targeted follow-up diagnostics. This ledger separates these units so that engineering checks, hyperparameter trials, and human examples are not inflated into a common count of "experiments." No new training was run while preparing this manuscript revision.

I.1 Execution Order and Separation of Responsibilities

1. **Lock source identity.** Retain response/preference/attribute table hashes, official model revision, tokenizer/pooling implementation, and candidate provenance. Verify that the source split and response feature indices agree with the cached representation table.
2. **Audit human supervision.** Assemble a blind A/B packet, hide requested criteria and generator metadata, randomize orientation, and retain a private key. Restore numeric differences before scoring. Freeze the V2 outcomes before revising the task-relative correctness rule.
3. **Build the trainable subset.** Exclude previously reviewed candidates, their reverse families, and their prompts. Check four-cell balance, nested Silver100 membership, exact reverse strings,

unique prompts, and full-context review records. Strip selection scores and reasons from model inputs.

4. **Prepare features once.** Verify exact cache identity, extract only missing endpoints with the frozen backbone, and fit normalization on the allowed TRAIN population. Hash the feature manifest, vector files, code snapshot, and exclusions before fitting.
5. **Train comparable heads.** Use the same three-coordinate architecture, preference pairs, optimizer, seed list, and DEV preference-loss checkpoint rule for the four main methods. Dense attribute supervision is an explicit difference, not an unreported input to the preference-only model.
6. **Freeze and score.** Save checkpoints and predictions before inspecting TEST labels. Fit BT mapping on DEV and scales on TRAIN, freeze both, then calculate TEST metrics and paired prompt intervals. Record raw BT only as a second view, not another fit.
7. **Diagnose without rewriting TEST.** Replay checkpoints, check edit shortcuts and nullspaces, perform same-pipeline extraction, and run natural-contrast and separate-head follow-ups on DEV. Mark incomplete adaptation separately. Confirmation requires a new locked evaluation, not selection of a favorable exploratory branch.

I.2 Completed Components and Their Evidence Units

Table 12: Verified scope of the focal archive. Rows count different units and should not be summed into a claimed number of independent confirmations. Historical layer/pooling and TinyHuman diagnostics are additionally retained in Appendix F.

Component	Recorded units	Purpose and qualification
Synthetic exact anchors	$8 \times 5 = 40$ fits	Eight nested budgets, five initializations; shared generated dataset.
Synthetic off-target noise	$4 \times 5 = 20$ fits	Four nonzero noise levels at fixed budget 25 per criterion.
Synthetic controls	$5 + 1 + 1$ fits	Five orthogonality fits; shuffled and sign-flipped controls each use one seed.
Human audits	80 + 20 pairs	Historical diagnostic audit and fresh quota-stratified V2 blind audit; not 100 independent raters.
Silver endpoint preparation	200 requests	100 exact cache hits, 100 new vectors, zero truncated requests.
Main frozen-head comparison	$4 \times 5 = 20$ fits	Original TEST study; five raw-BT rows reuse these fits.
Same-pass diagnostic	10 new fits	Five all-family and five intervention-family-holdout fits; original-endpoint references are reused.
Natural-contrast follow-up	$5 \times 5 = 25$ fits	DEV comparison with common additional exclusions and separate natural-pair holdout.
Semantic ranking grid	$3 \times 5 = 15$ fits	Three L2 values; final setting selected by mean DEV ranking loss.
Broader frozen-feature search	100 + 3 settings	Seed-0 exploratory screen plus cross-prompt settings; includes reused/warm-started candidates.
LoRA adaptation	4 interrupted runs	No completed final composition endpoint; excluded from performance claims.

Synthetic off-target standard deviations are $0.05, 0.1, 0.2, 0.5$; the clean budget grid is $0, 1, 5, 10, 25, 50, 100, 250$. Intervention pools are generated once per training seed at the maximum budget, and each smaller budget uses a prefix for each criterion. A gain across nested budgets is not based on newly independent datasets. All 67 listed synthetic configurations have saved configurations and metrics; the ledger does not treat their controls as independent real-text evidence.

The later real-text screen consists of 32 tree settings, 48 linear settings, eight MLP settings, six structured settings, and six corrected ensemble settings. The three cross-prompt settings are separate. The selected linear seed-0 head appears again as a screening candidate, and several nonlinear

models warm-start it. Counts therefore describe documented search scope, not 103 unrelated discoveries. The natural-contrast, ranking, and screening branches all reuse DEV; independent confirmation remains missing.

I.3 Actual Main Checkpoints and Optimization Budgets

All main runs initialize the bias-free projection using the framework's linear-layer default and the global context vector from a zero-mean normal with SD 0.3, seeded with 0–4. The training program uses full-batch AdamW, learning rate 0.003, weight decay 0.01, and gradient-norm clipping at 10. It evaluates after updates at step indices $0, 10, 20, \ldots$ and at the last permitted step. A DEV loss improvement must exceed 10^{-5} to replace the saved state. Stopping is allowed from index 100 once 100 step indices have elapsed since the best checkpoint. This rule explains why a run ending at index 110 records 111 optimizer updates.

Table 13: Executed main training: selected checkpoint index / optimizer updates. The maximum allowed budget is 600, but the completed main runs stop earlier. Indices are zero-based; updates count actual iterations.

Method	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4
Aligned BT	20/121	10/111	10/111	10/111	10/111
Silver-200	10/111	30/131	30/131	20/121	30/131
Silver-100	10/111	10/111	30/131	20/121	30/131
Dense attribute	20/121	30/131	10/111	20/121	10/111

The selected state is cloned when DEV loss improves and restored before producing predictions. TEST labels never choose the checkpoint. Each completed run stores a state file, predictions (including feature mean/scale and response identities), a learning curve, and a completion record containing their hashes. A later run continued to index 599 is an explicit post-study diagnostic, not the state evaluated in the frozen TEST table.

The separated-ranking branch retains the same learning rate, optimizer weight decay, clipping, and evaluation schedule but selects checkpoints by the equal-axis DEV ranking BCE. The explicit L2 grid is $0, 0.1, 1$ in addition to AdamW decay. Across its five fits per setting, the mean selected DEV ranking losses are 0.60192, 0.54563, and 0.56224; thus 0.1 is selected without directly maximizing composition. The corresponding DEV compositions are 71.78%, 72.67%, and 70.91%. This distinction describes the recorded selection rule but does not eliminate selection uncertainty or prior DEV reuse.

I.4 Auditable Artifacts and Source-to-Claim Links

For the completed two-axis update, paths below are relative to `experiments/clean2d_updates/20260925T182503Z/payload/` in the experiment archive.

Training protocol and source freeze
`configs/clean2d_v21/experiment_protocol.json` and `configs/clean2d_v21/source_lock.json`. The code snapshot contains the executed head architecture, feature extractor, and training loop. Configuration metadata alone is insufficient when an option is not consumed by the loop.

Exclusions and cache validation
`outputs/clean2d_v21/complete_data_validation.json`, `prepared_training_identity.json` and `cache_preflight.json` in the same output directory. These record common TRAIN identities, exact endpoint reuse and overlap checks. Private annotation identities are not reproduced in this manuscript package.

Completed fits and predictions
`outputs/clean2d_v21/runs/` contains the four method directories and five seeds each. Completion records distinguish actual finished fits from planned budgets. No raw-BT diagnostic row creates another checkpoint.

Evaluation and resampling
`scripts/clean2d_v21_eval/evaluate.py` and `outputs/clean2d_v21/evaluation/`. Saved mappings, evaluation arrays, prompt bootstrap counts, distributions and summaries make the reported uncertainty recomputable without fitting again.

Feature and nullspace diagnostics
`scripts/clean2d_v21_diagnostics/` and `outputs/clean2d_v21/diagnostics/`. These distinguish same-pass extraction, family holdout, template-compatible rules and reward-preserving feature-nullspace perturbations.

Small-head recovery and further search
`scripts/clean2d_v22/run.py`, `scripts/clean2d_v23/rank.py`, `outputs/clean2d_v23/DEV_paired_bootstrap.json`, and `scripts/clean2d_v24/`. The ranking objective and fixed residual construction govern the 10,240-parameter claim.

The latest snapshot manifest contains 1,395 payload entries; byte counts and SHA256 values were checked against the local archive. A matching digest verifies file identity, not the scientific validity of every assumption. The earlier Silver preparation record's "training not executed" flag describes that older release, while the later completion records establish the 20 fitted main models. Neither status should overwrite the other's historical meaning.

I.5 What the Paper Package Can Reproduce Locally

The source package includes the LaTeX manuscript, aggregate chart CSVs, figure assets, the parameter-accounting checks, and the displayed implementation excerpts. `code/check_appendix_math.py` checks algebraic invariants, analytic gradients against finite differences, reverse-pair equivalence, and bootstrap aggregation on deterministic toy arrays using NumPy. These tests do not execute model training or add new empirical results. The PyTorch extraction/training excerpts document the mechanism but require the original model/data environment to run end to end.

With the sibling experiment archive available, `verify_paper.py` recomputes manuscript aggregates, checks source/asset hashes and references, and enforces the nine-page main-text limit. `verify_frozen_appendix.py` checks the frozen architecture, parameter counts, completion ledger, selected ranking results, and code/configuration distinction. The complete training cache and private blind keys are not embedded in the Overleaf package. Reproduction of the original large-model feature extraction therefore requires the separately archived data and revision-locked backbone; compiling the paper does not.

J Outstanding Empirical Questions

The matched five-seed Silver comparison is now complete and negative on the original TEST; it must not be replaced by selectively reported DEV improvements. The main remaining test is an independently evaluated, frozen semantic-head candidate with matched supervision controls. Broader intervention families and cross-construction holdouts are needed to test whether coverage, rather than another correlated design choice, explains the failure. V2.1 still requires a fresh human audit because the V2 set informed its task-relative guard. Multiple independent annotators are needed to assess reliability. A semantic decomposition that contributes nonzero terms to reward while retaining preference utility remains unestablished; the separated-head construction does not answer that question. The interrupted adaptation runs provide no final evidence on whether representation learning repairs the deficit.