Beyond Semantic Alignment: Geometric Incomparability in Multi-Oracle Soft Fusion

Xiaodong Li[1], Binsheng Zhao[1], Yan Chen[2], Bingbing Jiang[3], Feijiang Li[1], Peng Zhou[4], Yuhua Qian[1], Liang Du[1,*]
[1]Shanxi University
{lixiaodong2, zhaobinsheng, fjli, duliang}@sxu.edu.cn, jinchengqyh@163.com
[2]Taiyuan University of Technology
chenyan01@tyut.edu.cn
[3]Hangzhou Normal University
jiangbb@hznu.edu.cn
[4]Anhui University
pengzhou@ahu.edu.cn

Abstract

Many decision-level fusion methods implicitly assume that soft outputs from different oracles can be directly aggregated once they are semantically aligned on the same label simplex. However, this assumption often breaks down in practice. Oracle-specific discrepancies in confidence scaling and boundary uncertainty may drive soft outputs into disparate geometric manifolds, rendering direct fusion inherently biased despite identical label semantics. To address this issue, we formalize this failure mode as geometric incomparability and propose a pre-fusion comparability recovery framework that learns a shared consensus via constrained oracle-specific rectifiers on the probability simplex. The framework includes a closed-form variant with frozen geometry descriptors and an alternating-refinement variant with consensus-dependent descriptors. Theoretically, we characterize the rectifier family, establish exact recovery under matched radial distortions, characterize the bias of direct fusion, and establish stable recovery under approximate mismatch. Experiments on controlled synthetic stress tests verify when direct fusion fails and when comparability recovery succeeds. Furthermore, real black-box evaluations demonstrate that similar geometric heterogeneity persists in practical multi-oracle scenarios. These results indicate that restoring probability-level comparability is an important step toward reliable soft-output fusion. The code is available at https://github.com/Lixiaodong-ml/CFCR-ACR.

1 Introduction

Decision-level fusion is a foundational design pattern for integrating probabilistic outputs from multiple sources, including fuzzy clusterer ensembles [1, 2], classifier ensembles [3], expert aggregation [4], soft crowdsourcing [5], and collaborative systems built on large language models (LLMs) [6]. Given multiple posterior vectors in the same answer space, a typical strategy is to average, weight, calibrate, or otherwise aggregate them to obtain a consensus prediction [7, 8].

However, most existing fusion methods implicitly assume that semantic alignment and a shared output simplex are sufficient to make oracle outputs directly comparable for fusion [9, 10]. In this paper, we question this premise. Semantic alignment does not necessarily imply geometric comparability. Even

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

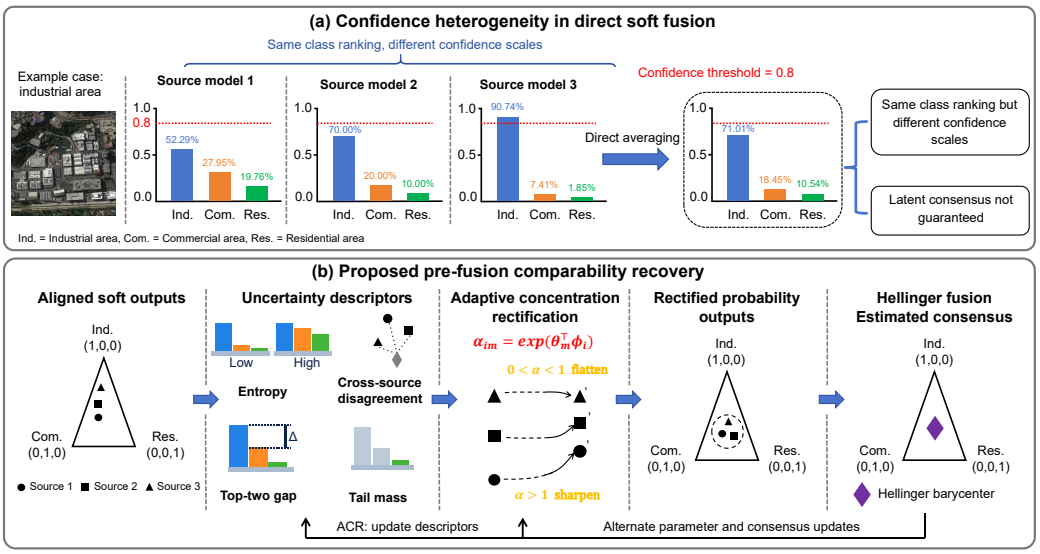


Figure 1: Illustration of geometric incomparability and comparability recovery. (a) shows how geometric mismatch biases direct averaging despite semantic alignment, distorting confidence and uncertainty. (b) shows the recovery process, where geometric descriptors calibrate posterior concentration on the simplex to preserve rank structure and promote geometric alignment before fusion.

when different oracles refer to the same classes, they may encode confidence via different posterior geometries, such as over-sharp concentration, excessive uncertainty spread, boundary ambiguity, or distorted tail allocation. As a result, directly averaging individually meaningful posteriors can still yield a biased consensus. We refer to this failure mode as geometric incomparability. It describes the situation where aligned soft outputs can be viewed as oracle-specific geometric distortions of a common latent consensus distribution, yet direct fusion of the raw outputs does not reliably recover that consensus. Unlike single-model calibration, supervised stacking, or discrete truth inference, our goal is label-free comparability recovery among multiple simplex-valued outputs before fusion, without requiring validation labels or access to oracle internals. An intuitive overview of this failure mode and the subsequent recovery process is given in Figure 1.

To this end, we propose a pre-fusion comparability recovery framework. Rather than designing another weighting rule on raw outputs, we rectify each oracle posterior on the probability simplex via a learned order-respecting transformation, followed by aggregation through a Hellinger barycenter. The rectifier mitigates geometric incomparability by adjusting posterior concentration while preserving simplex structure and semantic order. We instantiate this framework with two variants: a frozen-descriptor version with closed-form barycenter updates and a consensus-dependent alternating-refinement version. We further delineate the feasible regime for comparability recovery and analyze the impact of approximate mismatch on the recovered consensus. The main contributions of this paper are:

- We identify posterior-geometry mismatch as a source of bias in semantically aligned soft fusion and term it geometric incomparability. To address this, we propose CFCR and ACR methods to rectify oracle-specific probability geometry before Hellinger barycenter fusion.
- We theoretically characterize the proposed rectifier by first revealing the bias in direct fusion, then deriving the conditions for exact recovery under matched scalar radial distortions, and finally proving stability under approximate mismatch.
- Synthetic stress tests demonstrate that direct fusion is biased under probabilistic geometric mismatch, while real-world black-box results further corroborate this phenomenon.

2 Problem Setting: Geometric Comparability in Aligned Soft Fusion

Rather than asking how aligned oracle outputs should be weighted after they are produced, we ask a more basic question: whether these outputs are geometrically comparable before fusion. This

question is distinct from semantic alignment, calibration of a single model, and discrete truth inference [11, 12]. It concerns the geometry of multiple simplex-valued outputs that already share the same coarse label semantics. Formally, for each instance $i \in \{1, \dots, n\}$ and each oracle $m \in \{1, \dots, M\}$, we observe a soft output

$$\mathbf{u}_i^{(m)} = \big(u_{i1}^{(m)}, \dots, u_{iK}^{(m)}\big) \in \Delta^{K-1},$$

where Δ^{K-1} denotes the probability simplex over K classes. We focus on the aligned soft-fusion setting, where the k-th component has the same coarse semantic meaning across all oracles, and all outputs lie on the same simplex. Thus, the difficulty considered here is not label permutation, inconsistent output support, or component-level semantic rewriting.

Even under this aligned setting, oracle outputs may differ in posterior geometry. Some oracles may produce sharper distributions, others may be more diffuse, and some may allocate probability mass differently to tail classes or boundary regions. These differences can make direct aggregation biased despite semantic compatibility. To formalize this regime, we posit that each instance i has an oracle-independent latent consensus distribution $\boldsymbol{\pi}_i \in \Delta^{K-1}$, which denotes the shared latent posterior behind oracle-specific geometric distortions.

Definition 1 (Geometric Incomparability) *A collection of aligned soft outputs $\{\mathbf{u}_i^{(m)}\}_{m=1}^M$ is said to exhibit geometric incomparability if the outputs can be regarded as oracle-specific geometric distortions of a latent consensus distribution $\boldsymbol{\pi}_i \in \Delta^{K-1}$, while direct fusion of the raw outputs may yield a consensus biased away from $\boldsymbol{\pi}_i$.*

This definition formalizes the key phenomenon considered in this paper, where shared label semantics alone do not guarantee posterior-geometry comparability. Accordingly, our goal is not simply to aggregate aligned soft outputs. Instead, for each instance i, we seek to recover a consensus distribution $\mathbf{q}_i \in \Delta^{K-1}$ that better approximates the latent target $\boldsymbol{\pi}_i$ than direct fusion of the raw oracle outputs. This motivates a pre-fusion comparability recovery mechanism that corrects oracle-specific geometric distortions while preserving simplex structure and coarse semantic ordering.

3 Methodology

3.1 Geometry-Preserving Simplex Rectification

To address these issues, we seek a pre-fusion rectification mechanism that improves geometric comparability across oracle outputs before aggregation. A natural but overly flexible design would be to assign a separate parameter to each component. However, such a parameterization may change the relative order of components and thereby blur the distinction between geometric correction and semantic rewriting. To avoid altering class-wise semantics, we adopt a deliberately restricted but expressive scalar rectifier, assigning a single exponent to each oracle-instance pair and applying it uniformly across all components. This design captures radial concentration mismatch while preserving simplex structure and component ordering, making the recovery analysis tractable.

For $\alpha > 0$, we define the rectifier

$$R_{\alpha,\varepsilon}(\mathbf{u})_k = \frac{(u_k + \varepsilon)^\alpha}{\sum_{j=1}^K (u_j + \varepsilon)^\alpha}, \qquad k = 1, \dots, K, \tag{1}$$

where $\varepsilon > 0$ is a small smoothing constant used for numerical stability. When $\alpha > 1$, the transformation sharpens the posterior; when $\alpha < 1$, it flattens it. Since the same exponent is applied to every component, the rectifier preserves coarse ordering while adjusting the concentration profile. For oracle m and instance i, the rectified output and its parameterized rectification strength are given by

$$\tilde{\mathbf{u}}_i^{(m)} = R_{\alpha_{im},\varepsilon}\big(\mathbf{u}_i^{(m)}\big), \qquad \alpha_{im} = \exp\big(\boldsymbol{\theta}_m^\top \boldsymbol{\phi}_i\big), \tag{2}$$

where $\boldsymbol{\theta}_m$ is an oracle-specific parameter vector and $\boldsymbol{\phi}_i$ is an instance-level geometry descriptor. In this way, the model allows different oracles to exhibit different geometric tendencies, while still adapting the correction strength to the local uncertainty profile of each instance. Unlike global temperature [11] or power correction [11, 13], α_{im} is oracle- and instance-dependent, enabling pre-fusion comparability recovery without altering semantic ordering.

3.2 Geometry Descriptors

Designing ϕ_i is crucial because it determines the strength of the correction each oracle applies to each instance. We consider two construction methods for this purpose.

Frozen Geometry Descriptors. The fixed descriptor ϕ_i is computed from the observed oracle outputs, yielding a frozen-descriptor variant of the comparability recovery procedure:

$$\phi_i^{\mathrm{frz}} = \left[1,\ \frac{1}{M}\sum_{m=1}^{M} ent\big(\mathbf{u}_i^{(m)}\big),\ \frac{1}{M}\sum_{m=1}^{M} H^2\big(\mathbf{u}_i^{(m)}, \bar{\mathbf{u}}_i\big),\ \frac{1}{M}\sum_{m=1}^{M} tail\big(\mathbf{u}_i^{(m)}\big)\right]^\top, \tag{3}$$

where $\bar{\mathbf{u}}_i = \frac{1}{M}\sum_{m=1}^{M}\mathbf{u}_i^{(m)}$, $ent(\mathbf{u}_i^{(m)}) = -\frac{1}{\log K}\sum_{k=1}^{K} u_{ik}^{(m)} \log u_{ik}^{(m)}$, $tail(\mathbf{u}_i^{(m)}) = 1 - u_{i,(1)}^{(m)} - u_{i,(2)}^{(m)}$, and $H(\cdot,\cdot)$ denotes the distance metric. Here, the first component is a bias term, the entropy term captures average uncertainty [14], the mean deviation from $\bar{\mathbf{u}}_i$ measures cross-oracle geometric dispersion, and the tail term reflects posterior mass outside the top two components $u_{i,(1)}$ and $u_{i,(2)}$.

Consensus-Refined Geometry Descriptors. The refined descriptor is updated from the current consensus, enabling adaptive rectification in the alternating-refinement scheme. Specifically, at iteration t, it is defined as

$$\phi_i^{(t)} = \left[1,\ 1 - \big(q_{i,(1)}^{(t)} - q_{i,(2)}^{(t)}\big),\ ent\big(\mathbf{q}_i^{(t)}\big),\ tail\big(\mathbf{q}_i^{(t)}\big),\ \frac{1}{M}\sum_{m=1}^{M} H^2\big(\mathbf{u}_i^{(m)}, \bar{\mathbf{u}}_i\big)\right]^\top. \tag{4}$$

Compared with the frozen version, this descriptor incorporates consensus-dependent information, including the tail mass outside the top-two classes, entropy, and tail behavior, while retaining an observed-output term for inter-oracle dispersion.

3.3 Hellinger-Based Comparability Recovery

Since both rectified oracle outputs and the consensus lie on the simplex, their discrepancy should respect simplex geometry and support efficient updates. We therefore build the framework on the Hellinger [15] distance:

$$H^2(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\sqrt{\mathbf{p}} - \sqrt{\mathbf{q}}\|_2^2, \tag{5}$$

where $\mathbf{p}$ and $\mathbf{q}$ are two vectors of the same dimension. Under the square-root map, the Hellinger distance becomes Euclidean distance in the positive orthant of the unit sphere, so the Hellinger barycenter admits a closed-form update once the rectified outputs are fixed [16]. We therefore jointly learn the consensus distributions and rectifier parameters by minimizing the aggregate Hellinger discrepancy

$$\min_{\mathbf{Q},\boldsymbol{\theta}} F(\mathbf{Q}, \boldsymbol{\theta}) := \sum_{i=1}^{n}\sum_{m=1}^{M} H^2\big(R_{\alpha_{im},\varepsilon}(\mathbf{u}_i^{(m)}), \mathbf{q}_i\big), \tag{6}$$

where $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_n]^\top \in \mathbb{R}^{n\times K}$, $\mathbf{q}_i \in \Delta^{K-1}$, and $\alpha_{im} = \exp(\boldsymbol{\theta}_m^\top \phi_i)$. To prevent all oracles from collectively drifting in the same geometric direction, we impose the following constraints on Eq. (6)

$$\theta_{md} \in [-B, B], \qquad \sum_{m=1}^{M}\theta_{md} = 0, \qquad d = 1, \ldots, D, \tag{7}$$

where D denotes the dimension of the geometry descriptor. The box constraint bounds the rectification strength, while the centering constraint removes redundant global drift across oracles.

3.4 Optimization and Complexity

When using Eq. (3) as the descriptor, the Closed Comparability Recovery (CFCR) method is obtained.

Update $\boldsymbol{\theta}$. With $\mathbf{Q}$ fixed, the parameter vector of each oracle can be updated by solving

$$\min_{\boldsymbol{\theta}_m \in [-B,B]^D} \sum_{i=1}^{n} H^2\big(R_{\exp(\boldsymbol{\theta}_m^\top \phi_i),\varepsilon}(\mathbf{u}_i^{(m)}), \mathbf{q}_i\big), \qquad m = 1, \ldots, M, \tag{8}$$

followed by projection onto the centering constraint [17].

Update Q. With $\boldsymbol{\Theta}$ fixed, the consensus update decomposes over instances. Specifically, the rectification strength and the corresponding corrected output are computed as

$$\alpha_{im}^{(t+1)} = \exp\big((\boldsymbol{\theta}_m^{(t+1)})^\top \phi_i\big), \qquad \tilde{\mathbf{u}}_i^{(m,t+1)} = R_{\alpha_{im}^{(t+1)},\varepsilon}\big(\mathbf{u}_i^{(m)}\big). \tag{9}$$

The consensus distribution is then updated from the current rectified outputs $\tilde{\mathbf{u}}_i^{(m,t+1)}$ by solving

$$\mathbf{q}_i^{(t+1)} = \arg\min_{\mathbf{q}\in\Delta^{K-1}} \sum_{m=1}^{M} H^2\big(\tilde{\mathbf{u}}_i^{(m,t+1)}, \mathbf{q}\big) \tag{10}$$

By the closed-form barycenter result in the Appendix, Eq. (10) has the closed-form solution

$$q_{ik}^{(t+1)} = \frac{\left(\sum_{m=1}^{M} \sqrt{\tilde{u}_{ik}^{(m,t+1)}}\right)^2}{\sum_{j=1}^{K}\left(\sum_{m=1}^{M} \sqrt{\tilde{u}_{ij}^{(m,t+1)}}\right)^2}, \tag{11}$$

where $k = 1, \ldots, K$. It should be emphasized that the update of $\mathbf{q}_i^{(t+1)}$ is performed on the rectified outputs rather than the raw oracle outputs. Hence, once comparability is restored, the resulting consensus coincides with the Hellinger barycenter of the corrected posteriors.

When using the consensus refinement descriptor in Eq. (4), we obtain the Alternate Comparability Refinement (ACR) method. Since the ACR update follows the same alternating structure, we summarize its update process in Algorithm 1 in the Appendix.

Complexity Analysis. In each iteration, rectification and barycenter updates cost $\mathcal{O}(nMK)$, while oracle-parameter updates cost $\mathcal{O}(nM(K+D))$ per projected gradient sweep. Since D is small and the method uses only output-space information, the overall per-iteration complexity is linear in the number of observed posterior entries, up to a low-dimensional parameter-update overhead.

4 Theoretical Analysis of Comparability Recovery

This section analyzes the recoverable regime of the proposed comparability recovery framework. Instead of claiming universal correction under arbitrary probability mismatch, we focus on scalar radial distortions and establish a chain of results.

4.1 Preliminaries

For theoretical clarity, we consider the idealized rectifier with $\varepsilon = 0$ on the interior of the simplex:

$$R_\alpha(\mathbf{u})_k = \frac{u_k^\alpha}{\sum_{j=1}^{K} u_j^\alpha}, \qquad \alpha > 0, \tag{12}$$

for $\mathbf{u} \in \Delta^{K-1}$ with strictly positive components. The smoothed version with $\varepsilon > 0$ is used only in implementation. The idealized form is adopted because it admits an exact multiplicative composition law, which is central to the recovery analysis below. Then, we introduce the theoretical setting under which the main results are established. The first assumption is the technical regularity condition, while the second assumption specifically explains the matching distortion model and its feasibility under the proposed descriptor parameterization.

Assumption 1 (Interior Simplex) *There exists a constant $\delta \in (0, 1/K)$ such that all latent consensus distributions considered in the theory satisfy $\pi_{ik} \geq \delta, \forall i, k$.*

Assumption 2 (Matched Distortion Setting) *For each instance i, there exists a latent consensus distribution $\boldsymbol{\pi}_i \in \Delta^{K-1}$ such that, for every oracle m, the following conditions hold:*

(i) the true rectification strength is representable by the descriptor model, i.e., there exists an oracle-specific vector $\boldsymbol{\theta}_m^\star$ satisfying $\alpha_{im}^\star = \exp\big((\boldsymbol{\theta}_m^\star)^\top \boldsymbol{\phi}_i\big)$;

(ii) the observed output $\mathbf{u}_i^{(m)}$ is obtained by applying an oracle-specific inverse radial distortion to the latent consensus $\boldsymbol{\pi}_i$, i.e., $\mathbf{u}_i^{(m)} = R_{1/\alpha_{im}^\star}(\boldsymbol{\pi}_i)$ with $\alpha_{im}^\star > 0$;

(iii) both the true rectification intensity $\alpha_{im}^\star$ and its estimate $\hat{\alpha}_{im}$ are confined within a bounded interval, i.e., there exist constants $0 < a^- < a^+ < \infty$ such that

$$a^- \le \alpha_{im}^\star \le a^+, \qquad a^- \le \hat{\alpha}_{im} \le a^+, \qquad \forall i, m.$$

4.2 Recovery Guarantees

Before establishing positive recovery guarantees, it is instructive to understand why direct fusion can fail even under matched distortions. To clarify this issue, we begin with the matched-distortion setting and examine the simplest nontrivial case, namely binary outputs. This setting is analytically transparent and already reveals the essential mechanism by which direct fusion becomes biased. Specifically, let $K = 2$ and write the latent consensus as $\boldsymbol{\pi}_i = (p_i, 1 - p_i), p_i \in (0, 1)$. Under the idealized rectifier family, the first component of the transformed distribution satisfies

$$(R_\alpha(\boldsymbol{\pi}_i))_1 = \frac{p_i^\alpha}{p_i^\alpha + (1 - p_i)^\alpha} = \sigma\big(\alpha \log \frac{p_i}{1 - p_i}\big), \tag{13}$$

where $\sigma(z) = 1/(1 + e^{-z})$. In the binary case, radial distortion therefore scales the latent logit multiplicatively, enabling explicit analysis of how oracle-specific distortions bias naive averaging.

Theorem 1 (Bias of Direct Averaging under Matched Distortions) *Under the matched distortion model, unequal nontrivial distortion strengths generally bias the direct average, yielding $\bar{u}_{i1} := \frac{1}{M}\sum_{m=1}^M u_{i1}^{(m)} \neq p_i$. Moreover, if $\eta_{im} = \alpha_{im}^{\star -1} - 1$ and $|\eta_{im}| \le \eta_{\max}$ with $\eta_{\max}$ sufficiently small,*

$$\bar{u}_{i1} - p_i = \sigma'(\ell_i)\ell_i\,\bar{\eta}_i + \frac{1}{2}\sigma''(\ell_i)\ell_i^2\,\overline{\eta_i^2} + O(\eta_{\max}^3), \tag{14}$$

where $\ell_i = \log \frac{p_i}{1-p_i}$, $\bar{\eta}_i = M^{-1}\sum_m \eta_{im}$, and $\overline{\eta_i^2} = M^{-1}\sum_m \eta_{im}^2$.

As established in Theorem 1, direct averaging is generally biased even when all oracle outputs are generated from the same latent consensus through matched radial distortions. Thus, semantic alignment and a shared simplex do not by themselves guarantee recoverability by raw aggregation.

Theorem 2 (Exact Recovery under Matched Scalar Radial Distortions) *Under Assumption 2, if the true rectification strengths $\alpha_{im}^\star$ are used, then for every instance i and oracle m,*

$$R_{\alpha_{im}^\star}(\mathbf{u}_i^{(m)}) = \boldsymbol{\pi}_i.$$

Consequently, the Hellinger barycenter of the rectified outputs satisfies $\mathbf{q}_i = \boldsymbol{\pi}_i$.

However, the exact recovery in Theorem 2 relies on a perfectly matched distortion model. To assess robustness beyond this ideal case, we consider the following approximate-mismatch condition.

Assumption 3 (Approximate Matched Distortion) *There exist latent consensus distributions $\boldsymbol{\pi}_i$ and ideal rectification strengths $\tilde{\alpha}_{im}$ such that, for all i, m,*

$$H\big(\mathbf{u}_i^{(m)}, R_{1/\tilde{\alpha}_{im}}(\boldsymbol{\pi}_i)\big) \le \xi_{im}, \qquad |\hat{\alpha}_{im} - \tilde{\alpha}_{im}| \le \Delta_{im}.$$

Under this condition, the stability of the proposed framework depends on controlling three error-propagation mechanisms: mismatch in the observed outputs, quantified by ξ_{im}; estimation error in the rectification strengths, quantified by Δ_{im}; and the stability of the Hellinger barycenter map that determines the final consensus. We discuss them sequentially in Lemmas 1–3.

Lemma 1 (Lipschitz Continuity in the Exponent) *For any $c \in (0, 1/K)$, there exists a constant $L = L(c, a^-, a^+) < \infty$ such that for all $\mathbf{u} \in S_c$ and all $\alpha, \beta \in [a^-, a^+]$,*

$$H\big(R_\alpha(\mathbf{u}), R_\beta(\mathbf{u})\big) \le L|\alpha - \beta|,$$

where $S_c := \{\mathbf{u} \in \Delta^{K-1} : u_k \ge c,\ \forall k\}$.

Lemma 2 (Local Lipschitz Continuity in the Input) *For any $c \in (0, 1/K)$, there exists a constant $\hat{L} = \hat{L}(c, a^-, a^+) < \infty$ such that for all $\mathbf{u}, \mathbf{v} \in S_c$ and all $\alpha \in [a^-, a^+]$,*

$$H\big(R_\alpha(\mathbf{u}), R_\alpha(\mathbf{v})\big) \le \hat{L} H(\mathbf{u}, \mathbf{v}).$$

Lemma 3 (Barycenter Stability) *Let $\{p_m\}_{m=1}^M$ and $\{q_m\}_{m=1}^M$ be two collections in Δ^{K-1}, and let $\bar{p}$ and $\bar{q}$ denote their Hellinger barycenters. Then, we have*

$$H(\bar{p}, \bar{q}) \le 2\sqrt{K} \cdot \frac{1}{M} \sum_{m=1}^{M} H(p_m, q_m).$$

Combining the above Lemmas 1−3, the recovered consensus admits the following stability guarantee under approximate mismatch.

Theorem 3 (Stable Recovery under Approximate Mismatch) *Suppose Assumptions 1 and 3 hold, and both the observed oracle outputs and their ideal matched counterparts belong to a common interior simplex domain $\mathcal{S}_c$. Let $\mathbf{q}_i$ denote the Hellinger barycenter of the learned rectified outputs $\{R_{\hat{\alpha}_{im}}(\mathbf{u}_i^{(m)})\}_{m=1}^M$. Then there exist constants $C_1, C_2 > 0$, depending only on (δ, a^-, a^+, K), such that for every instance i,*

$$H(\mathbf{q}_i, \boldsymbol{\pi}_i) \le C_1 \cdot \frac{1}{M} \sum_{m=1}^{M} \xi_{im} + C_2 \cdot \frac{1}{M} \sum_{m=1}^{M} \Delta_{im}. \tag{15}$$

This result shows that the recovery error is linearly governed by the mismatch between the observed outputs and the matched radial model, together with the estimation error of the rectification strengths. Consequently, when both errors are small, the recovered consensus remains close to the latent target.

4.3 Scope of the Scalar Rectifier

The preceding theory also clarifies the scope of the proposed rectifier: its recoverability guarantees rely essentially on a scalar radial distortion model. We make the limitations of the proposed method explicit in the following proposition.

Proposition 1 (Limitation beyond Scalar Radial Distortions) *If oracle-specific distortions are component-anisotropic or fail to preserve coordinate ordering, a scalar family of rectification strengths α_{im} is generally insufficient to map all observed outputs back to a shared latent consensus.*

5 Experiments

In this section, we conduct mechanism-oriented experiments and scenario-level validations to examine the effectiveness of the proposed framework and its consistency with the theoretical analysis.

5.1 Datasets

Our evaluation involves two types of datasets. For mechanism-oriented experiments, we construct four synthetic datasets, denoted as D1–D4, by generating simplex-valued latent consensus distributions and applying controlled oracle-specific distortions. They allow direct examination of sharpness mismatch, boundary uncertainty, tail-mass inconsistency, and failures beyond the scalar radial distortion family. For scenario-level validation, we use real-world datasets from diverse domains, including AID [18], DomainNet [19], CIFAR-19 [20], Stanford-40-Actions [21], and two text toxicity datasets, HateXplain [22] and CivilComments [23]. Details of the datasets are provided in Table 1.

5.2 Experimental Setup

The experiments are designed to validate the proposed failure mode and the corresponding recovery mechanism. Synthetic experiments provide controlled access to the latent consensus, while real-world black-box validation examines whether similar probability-geometry heterogeneity arises in

Table 1: Statistics of the datasets used in the experiments.

Real-world datasets				Synthetic datasets			
Datasets	Instances	Classes	Oracles	Datasets	Instances	Classes	Oracles
Stanford-40-Actions	9,532	40	5	D1	1,000	8	10
AID	10,000	30	5	D2	5,000	13	7
CIFAR-19	10,000	19	5	D3	10,000	60	5
DomainNet	13,599	25	5	D4	30,000	40	12
HateXplain	19,229	3	3	–	–	–	–
CivilComments	45,180	2	3	–	–	–	–

practical multi-oracle outputs. To isolate the role of comparability recovery, CFCR and ACR are compared with arithmetic averaging (ME) [24], normalized geometric averaging (GeoM) [25], and unrectified Hellinger barycenter fusion (HRaw) [26]. We further include SCPP [27] and FMSE [9] as representative fuzzy clustering ensemble methods to provide broader comparisons under the label-free setting.

Mechanism-oriented experiments. We design six controlled experiments to examine the behavior and scope of comparability recovery. They test direct-fusion bias under matched distortions, boundary sensitivity through top-two-gap splits, tail-mass inconsistency, geometry diagnostics before and after rectification, descriptor ablations, and negative controls with coordinate-anisotropic or non-order-preserving distortions. These experiments jointly evaluate when the scalar rectifier is effective and where its theoretical limits appear.

Scenario-level validation. For each real-world dataset, Qwen [28], LLaMA [29], Microsoft [30], Hugging Face [31], and OpenBMB [32] are queried with identical prompts and candidate labels. The resulting class-wise scores are normalized into probability vectors, ensuring that all oracle outputs are semantically aligned and defined on the same label simplex. However, semantic alignment alone does not guarantee geometric comparability, since different models may exhibit distinct calibration behaviors, confidence sharpness, and uncertainty allocation patterns. Thus, this setting provides supporting evidence for whether probability-geometry heterogeneity also appears in practical black-box outputs and whether pre-fusion rectification can improve direct aggregation.

Evaluation metrics are selected to match the reference signal available in each setting. For synthetic datasets, the latent consensus distribution $\boldsymbol{\pi}_i$ is known, which allows direct evaluation of distribution-level recovery. Therefore, we report the average squared Hellinger error and KL [33] error, defined as $\mathrm{HellErr} = \frac{1}{n}\sum_{i=1}^{n} H^2(\mathbf{q}_i, \boldsymbol{\pi}_i)$ and $\mathrm{KLErr} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \pi_{ik} \log \frac{\pi_{ik}}{q_{ik}}$. When analyzing robustness under varying distortion strengths, accuracy (ACC) is also reported using the synthetic label to evaluate the decision-level impact of distributional recovery errors. For real-world datasets, the latent consensus distribution is unobserved, so HellErr and KLErr are not applicable. We use ACC and Brier score [34] as the primary metrics, evaluating decision-level correctness and probabilistic prediction quality under observed labels, defined as $\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \left(q_{ik} - \mathbf{1}\{y_i = k\}\right)^2$.

5.3 Experimental Results

In all tables, the best results are highlighted in bold, and the second-best are underlined. For readability, HellErr, KLErr, and Brier scores are reported after multiplication by 100. The synthetic results in Table 2 show that CFCR and ACR substantially reduce HellErr and KLErr compared with ME, GeoM, and HRaw across matched geometric distortions, with particularly clear gains on boundary-sensitive samples and tail-mass splits. These results confirm that direct aggregation can remain biased even when oracle outputs are semantically aligned on a shared label simplex. ACR often achieves the lowest error rates, suggesting that consensus-refined descriptors provide a more adaptive estimate of oracle-specific comparability. The additional boundary and tail-sensitive analyses in the Appendix further show that the gains persist on low-gap and high-tail subsets, supporting the roles of boundary uncertainty and tail-mass allocation in geometric incomparability.

To verify whether rectification improves geometric comparability, we report oracle-level diagnostics before and after correction in Table 3. EntStd, GapStd, and TailStd measure cross-oracle dispersion in entropy, top-two gap, and tail mass, while PairH measures average pairwise Hellinger discrepancy. As shown, CFCR and ACR consistently reduce these values relative to raw outputs, indicating that

Table 2: Comparison of consensus recovery errors across fusion methods in semantically aligned but geometrically incomparable synthetic scenarios.

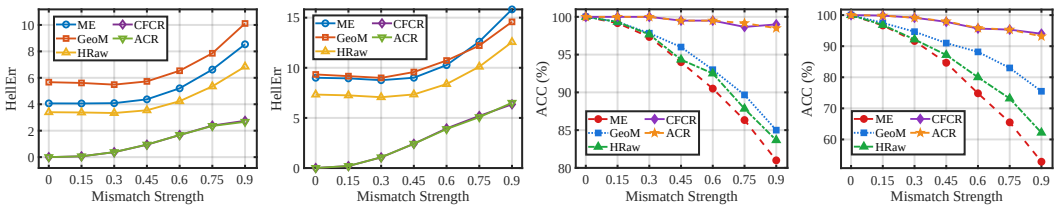
Datasets	HellErr ↓					KLErr ↓				
	ME	GeoM	HRaw	CFCR	ACR	ME	GeoM	HRaw	CFCR	ACR
D1_global	4.058	5.610	3.400	**0.000**	**0.000**	0.686	1.388	0.486	**0.000**	**0.000**
D1_boundary	2.170	3.436	1.743	0.798	**0.089**	0.354	1.618	0.212	0.061	**0.002**
D1_tail	2.464	3.444	1.981	0.221	**0.107**	0.464	1.266	0.287	0.006	**0.002**
D1_mixed	1.988	3.043	1.595	0.199	**0.041**	0.350	1.549	0.203	0.004	**0.002**
D2_global	5.247	6.824	4.429	**0.000**	**0.000**	1.131	2.030	0.817	**0.000**	**0.000**
D2_boundary	2.660	3.987	2.222	1.115	**0.247**	0.501	2.373	0.368	0.112	**0.010**
D2_tail	2.968	4.203	2.469	0.480	**0.316**	0.645	2.182	0.455	0.018	**0.009**
D2_mixed	2.401	3.551	2.006	0.333	**0.130**	0.483	2.275	0.348	0.010	**0.007**
D3_global	8.989	9.312	7.300	**0.006**	**0.006**	3.256	3.677	2.154	**0.000**	**0.000**
D3_boundary	4.164	4.893	3.450	1.700	**1.122**	1.312	4.264	1.032	0.255	**0.110**
D3_tail	4.599	5.574	3.828	1.617	**1.524**	1.560	4.680	1.202	0.157	**0.156**
D3_mixed	3.695	4.463	3.079	0.931	**0.906**	1.212	4.228	0.968	**0.065**	0.068
D4_global	6.722	7.046	5.446	**0.002**	**0.002**	1.848	2.107	1.218	**0.000**	**0.000**
D4_boundary	3.216	3.995	2.676	1.263	**0.608**	0.838	3.138	0.668	0.149	**0.041**
D4_tail	3.557	4.458	2.958	0.890	**0.837**	0.973	3.146	0.743	0.051	**0.049**
D4_mixed	2.881	3.654	2.405	0.520	**0.463**	0.783	3.059	0.625	0.025	**0.023**

Table 3: Oracle-level geometric diagnostics before and after rectification.

Data	EntStd ↓			GapStd ↓			TailStd ↓			PairH ↓		
	Raw	CFCR	ACR	Raw	CFCR	ACR	Raw	CFCR	ACR	Raw	CFCR	ACR
D1	11.320	3.910	**0.934**	9.068	2.898	**0.749**	10.237	3.730	**0.785**	12.151	4.107	**0.782**
D2	12.024	4.439	**2.211**	8.759	3.093	**1.599**	12.089	4.776	**2.324**	14.543	5.358	**2.391**
D3	11.609	5.296	**5.289**	6.374	2.551	**2.542**	11.782	5.691	**5.673**	18.139	9.050	**8.828**
D4	10.112	4.035	**3.752**	5.955	2.199	**2.008**	10.569	4.531	**4.193**	14.282	6.078	**5.409**

rectification decreases uncertainty, boundary, and tail-mass discrepancies before fusion. Descriptor ablation experiments in the appendix further confirm the roles of boundary, tail, and disagreement cues in probability-level comparability recovery.

The robustness curves in Figure 2 show that direct fusion methods become more sensitive as mismatch strength increases, whereas CFCR and ACR remain comparatively stable in both recovery error and decision accuracy. This trend is consistent with the theoretical prediction that rectifying oracle-specific probability geometry can reduce direct-fusion bias. For real black-box outputs, Tables 4 and 5 report ACC and Brier score under the same oracle pool. CFCR achieves the best average rank on both metrics, while ACR remains second-best on average, suggesting that pre-fusion comparability recovery can improve practical fusion quality. The contrast with synthetic results indicates that ACR benefits more from consensus refinement under recoverable distortions, whereas CFCR is more stable under complex black-box mismatches. Appendix diagnostics on real outputs show that black-box geometric incomparability arises from coupled uncertainty, boundary, tail, and model-specific mismatches rather than a single marginal factor. Although marginal diagnostics vary across datasets, PairH consistently decreases after rectification, suggesting improved probability-geometry alignment.



(a) D1-HellErr (b) D3-HellErr (c) D1-ACC (d) D3-ACC

Figure 2: Robustness comparison across methods on D1 and D3 under increasing mismatch strength.

Table 4: ACC comparison of different methods on real-world black-box outputs, where higher values indicate better performance and "OOM" denotes out-of-memory failure.

Datasets	ME	GeoM	HRaw	SCPP-a	SCPP-v	FMSE-a	FMSE-v	CFCR	ACR
AID	**61.150**	55.820	60.240	40.192	34.941	31.290	31.253	60.320	60.400
HateXplain	42.623	41.630	42.358	40.681	43.082	40.660	40.660	**43.258**	42.920
CIFAR-19	12.740	10.080	12.220	**18.58**0	15.473	10.020	10.020	14.520	14.550
CivilComments	74.358	72.913	74.154	OOM	OOM	OOM	OOM	**74.969**	74.398
DomainNet	74.601	71.976	74.733	54.112	55.261	51.670	51.643	**76.042**	74.991
Stanford-40-Actions	92.583	87.789	91.219	80.232	76.251	42.082	42.100	**92.604**	92.142
Avg Rank	3.17	5.50	4.33	5.40	4.80	8.40	8.60	**1.83**	2.50

Table 5: Brier score comparison of different methods on real-world black-box outputs, where lower values indicate better performance, "NaN" denotes that the corresponding method is not applicable to this evaluation metric, and "OOM" indicates an out-of-memory failure.

Datasets	ME	GeoM	HRaw	SCPP-a	SCPP-v	FMSE-a	FMSE-v	CFCR	ACR
AID	59.889	60.608	58.259	103.542	NaN	126.590	NaN	**57.907**	57.918
HateXplain	**84.231**	94.733	87.951	97.563	NaN	138.591	NaN	86.848	86.970
CIFAR-19	95.760	103.290	97.568	111.971	NaN	150.394	NaN	**94.451**	95.703
CivilComments	**40.010**	49.948	43.045	OOM	NaN	OOM	NaN	41.505	41.841
DomainNet	39.518	38.582	36.975	88.022	NaN	107.703	NaN	**36.550**	36.762
Stanford-40-Actions	39.528	35.687	36.473	86.192	NaN	97.662	NaN	**35.439**	35.765
Avg Rank	3.17	4.33	3.67	6.00	—	7.00	—	**1.33**	2.50

6 Conclusion

This paper identified geometric incomparability as a failure mode in aligned multi-oracle soft fusion and proposed CFCR and ACR for pre-fusion comparability recovery. Theoretical analysis and synthetic experiments demonstrate the potential of rectification to reduce direct-fusion bias, while real-world black-box evaluations support its practical effectiveness in improving fusion performance. Current guarantees focus on scalar radial distortions, while component-wise and non-order-preserving mismatches remain open for future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62376146 and U2541227.

References

[1] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[2] Musa Mojarad, Samad Nejatian, Hamid Parvin, and Majid Mohammadpoor. A fuzzy clustering ensemble based on cluster clustering and iterative fusion of base clusters: M. mojarad et al. *Applied Intelligence*, 49(7):2567–2581, 2019.

[3] Sergey Tulyakov, Stefan Jaeger, Venu Govindaraju, and David Doermann. Review of classifier combination methods. *Machine learning in document analysis and recognition*, pages 361–386, 2008.

[4] Robert T Clemen and Robert L Winkler. *Aggregating probability distributions*. na, 2007.

[5] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and CK Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment (PVLDB)*, 2017.

[6] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[7] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 2002.

[8] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[9] Xia Ji, Jiawei Sun, Jianhua Peng, Yue Pang, and Peng Zhou. Clustering ensemble based on fuzzy matrix self-enhancement. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):148–161, 2024.

[10] Oscar Oelrich, Mattias Villani, and Sebastian Ankargren. Local prediction pools. *Journal of Forecasting*, 43(1):103–117, 2024.

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

[12] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C*, 28(1):20–28, 1979.

[13] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

[14] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[15] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[16] Frank Nielsen and Richard Nock. On the centroids of symmetrized bregman divergences. *arXiv preprint arXiv:0711.3242*, 2007.

[17] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

[18] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[19] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[21] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas J. Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1331–1338, 2011.

[22] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 14867–14875, 2021.

[23] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[24] Mervyn Stone. The opinion pool. *The Annals of Mathematical Statistics*, pages 1339–1342, 1961.

[25] Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

[26] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[27] Feijiang Li, Jieting Wang, Yuhua Qian, Guoqing Liu, and Keqi Wang. Fuzzy ensemble clustering based on self-coassociation and prototype propagation. *IEEE Transactions on Fuzzy Systems*, 31(10):3610–3623, 2023.

[28] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[29] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[30] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[31] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

[32] Guoyang Zeng, Xu Han, Zhengyan Zhang, Zhiyuan Liu, Yankai Lin, and Maosong Sun. Openbmb: Big model systems for large-scale representation learning. In *Representation Learning for Natural Language Processing*, pages 463–489. Springer, 2023.

[33] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[34] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Supplementary Material for "Beyond Semantic Alignment: Geometric Incomparability in Multi-Oracle Soft Fusion"

7 ACR Algorithm Update Process

Algorithm 1 Alternating Comparability Refinement (ACR).

Input: Soft outputs $\{\mathbf{u}_i^{(m)}\}$, bound B, and maximum iterations T
Output: Consensus matrix $\mathbf{Q}$ and oracle-parameter matrix $\boldsymbol{\theta}$

Initialize $\mathbf{Q}^{(0)}$ according to Eq. (11) and set $\boldsymbol{\theta}^{(0)} = \mathbf{0}$
repeat
  for $i = 1, \ldots, n$ **do**
    Compute the consensus-refined descriptor $\boldsymbol{\phi}_i^{(t)}$ according to Eq. (4)
  end for
  for $m = 1, \ldots, M$ **do**
    Update $\boldsymbol{\theta}_m^{(t+1)}$ by approximately solving Eq. (8)
  end for
  Project $\boldsymbol{\theta}^{(t+1)}$ onto the centering constraints in Eq. (7)
  for $i = 1, \ldots, n$ **do**
    for $m = 1, \ldots, M$ **do**
      Compute $\alpha_{im}^{(t+1)} = \exp\big((\boldsymbol{\theta}_m^{(t+1)})^\top \boldsymbol{\phi}_i^{(t)}\big)$
      Compute $\tilde{\mathbf{u}}_i^{(m,t+1)} = R_{\alpha_{im}^{(t+1)},\varepsilon}\big(\mathbf{u}_i^{(m)}\big)$
    end for
    Update $\mathbf{q}_i^{(t+1)}$ according to Eq. (11)
  end for
until convergence

8 Proof of Lemma 4.

Lemma 4 (Closed Form of Hellinger Barycenter) *Given distributions $p_1, \ldots, p_M \in \Delta^{K-1}$, the unique minimizer of $\min_{\mathbf{q}\in\Delta^{K-1}} \sum_{m=1}^M H^2(\mathbf{p}_m, \mathbf{q})$ is given componentwise by*

$$q_k^\star = \frac{\Big(\sum_{m=1}^M \sqrt{p_{mk}}\Big)^2}{\sum_{j=1}^K \Big(\sum_{m=1}^M \sqrt{p_{mj}}\Big)^2}, \qquad k = 1, \ldots, K.$$

Proof 1 *Because*

$$H^2(\mathbf{p}_m, \mathbf{q}) = \frac{1}{2}\sum_{k=1}^K \big(\sqrt{p_{mk}} - \sqrt{q_k}\big)^2 = 1 - \sum_{k=1}^K \sqrt{p_{mk} q_k}, \tag{16}$$

minimizing the sum of Hellinger distances is equivalent to maximizing

$$\sum_{m=1}^M \sum_{k=1}^K \sqrt{p_{mk} q_k} = \sum_{k=1}^K a_k \sqrt{q_k}, \qquad a_k := \sum_{m=1}^M \sqrt{p_{mk}}.$$

Let $x_k = \sqrt{q_k}$. Then the constraints become $x_k \geq 0$ and $\sum_k x_k^2 = 1$. The problem is

$$\max_{\mathbf{x}\geq 0,\, \|\mathbf{x}\|_2 = 1} \mathbf{a}^\top \mathbf{x}, \tag{17}$$

whose maximizer is the normalized vector $\mathbf{x}^\star = \mathbf{a}/\|\mathbf{a}\|_2$ because $\mathbf{a} \geq 0$. Therefore

$$q_k^\star = \frac{a_k^2}{\sum_{j=1}^K a_j^2}, \tag{18}$$

which is exactly the stated formula. Substituting the definition of a_k yields

$$q_k^\star = \frac{\left(\sum_{m=1}^M \sqrt{p_{mk}}\right)^2}{\sum_{j=1}^K \left(\sum_{m=1}^M \sqrt{p_{mj}}\right)^2}, \qquad k = 1, \ldots, K. \tag{19}$$

Therefore, the proof is complete.

9 Several Elementary Properties of the Idealized Rectifier.

We establish several elementary properties of the idealized rectifier. These results justify R_α as a valid simplex operator and provide the algebraic basis for the recovery analysis developed later.

Proposition 2 (Simplex Preservation) *For every $\alpha > 0$ and every $u \in \Delta^{K-1}$, we have*

$$R_\alpha(\mathbf{u}) \in \Delta^{K-1}.$$

Proof 2 *Each component of $R_\alpha(\mathbf{u})$ is nonnegative. Moreover,*

$$\sum_{k=1}^K R_\alpha(\mathbf{u})_k = \sum_{k=1}^K \frac{u_k^\alpha}{\sum_{j=1}^K u_j^\alpha} = \frac{\sum_{k=1}^K u_k^\alpha}{\sum_{j=1}^K u_j^\alpha} = 1. \tag{20}$$

Hence $R_\alpha(\mathbf{u})$ lies in Δ^{K-1}.

Proposition 3 (Order Preservation) *Let $\mathbf{u} \in \Delta^{K-1}$ and $\alpha > 0$. If $u_a \geq u_b$, then $(R_\alpha(\mathbf{u}))_a \geq (R_\alpha(\mathbf{u}))_b$, where $(R_\alpha(\mathbf{u}))_a$ represents the a-th component of $R_\alpha(\mathbf{u})$.*

Proof 3 *Since $x \mapsto x^\alpha$ is increasing on $[0, \infty)$ for every $\alpha > 0$, $u_a \geq u_b$ implies $u_a^\alpha \geq u_b^\alpha$, and division by the same positive denominator yields*

$$(R_\alpha(\mathbf{u}))_a \geq (R_\alpha(\mathbf{u}))_b. \tag{21}$$

Proposition 4 (Concentration Behavior) *For every $\mathbf{u} \in \Delta^{K-1}$, we have*

(i) $R_1(\mathbf{u}) = \mathbf{u}$;

(ii) if $\alpha > 1$, then $R_\alpha(\mathbf{u})$ is more concentrated than u in the sense that it increases the ratio between larger and smaller components;

(iii) if $\alpha < 1$, then $R_\alpha(\mathbf{u})$ is less concentrated than $\mathbf{u}$ in the same sense.

Proof 4 *Part (i) follows immediately from the definition of R_α. For (ii) and (iii), for any a, b with $u_a, u_b > 0$,*

$$\frac{(R_\alpha(\mathbf{u}))_a}{(R_\alpha(\mathbf{u}))_b} = \frac{u_a^\alpha}{u_b^\alpha} = \left(\frac{u_a}{u_b}\right)^\alpha. \tag{22}$$

If $u_a > u_b > 0$, this ratio is larger than u_a/u_b when $\alpha > 1$ and smaller than u_a/u_b when $\alpha < 1$.

Proposition 5 (Composition Law) *For all $\alpha, \beta > 0$ and all interior simplex vectors u, we have*

$$R_\alpha\big(R_\beta(\mathbf{u})\big) = R_{\alpha\beta}(\mathbf{u}).$$

Proof 5 *By definition,*

$$(R_\beta(\mathbf{u}))_k = \frac{u_k^\beta}{\sum_{j=1}^K u_j^\beta}. \tag{23}$$

Applying R_α again gives

$$(R_\alpha(R_\beta(\mathbf{u})))_k = \frac{\left(\frac{u_k^\beta}{\sum_j u_j^\beta}\right)^\alpha}{\sum_{\ell=1}^K \left(\frac{u_\ell^\beta}{\sum_j u_j^\beta}\right)^\alpha} = \frac{u_k^{\alpha\beta}}{\sum_{\ell=1}^K u_\ell^{\alpha\beta}} = (R_{\alpha\beta}(\mathbf{u}))_k, \tag{24}$$

which proves the claim.

10 Proof of Theorem 1.

Proof 6 *Under the matched distortion model, the first component of each oracle output can be written as*

$$u_{i1}^{(m)} = (R_{1/\alpha_{im}^\star}(\boldsymbol{\pi}_i))_1 = \frac{p_i^{\,1/\alpha_{im}^\star}}{p_i^{\,1/\alpha_{im}^\star} + (1-p_i)^{\,1/\alpha_{im}^\star}}. \tag{25}$$

Using the identity

$$\frac{p^\gamma}{p^\gamma + (1-p)^\gamma} = \sigma\left(\gamma \log \frac{p}{1-p}\right), \tag{26}$$

with $\gamma = 1/\alpha_{im}^\star$, we obtain

$$u_{i1}^{(m)} = \sigma\left(\frac{1}{\alpha_{im}^\star}\ell_i\right), \qquad \ell_i := \log \frac{p_i}{1-p_i}.$$

Then, we define $\eta_{im} := \alpha_{im}^{\star -1} - 1$, so that

$$\frac{1}{\alpha_{im}^\star} = 1 + \eta_{im}, \qquad u_{i1}^{(m)} = \sigma\big((1+\eta_{im})\ell_i\big).$$

Averaging over m gives

$$\bar{u}_{i1} = \frac{1}{M}\sum_{m=1}^{M} \sigma\big((1+\eta_{im})\ell_i\big). \tag{27}$$

We first prove the qualitative claim that, in general, $\bar{u}_{i1} \neq p_i$ whenever the distortion strengths are not all equal to 1. Since $p_i \neq \frac{1}{2}$, we have $\ell_i \neq 0$, and therefore the map

$$\eta \mapsto \sigma\big((1+\eta)\ell_i\big)$$

is nonconstant in a neighborhood of $\eta = 0$. Moreover,

$$\sigma(\ell_i) = p_i.$$

Thus, unless all $\eta_{im} = 0$ or a nongeneric cancellation happens across different m, the average

$$\frac{1}{M}\sum_{m=1}^{M} \sigma\big((1+\eta_{im})\ell_i\big) \tag{28}$$

does not equal $\sigma(\ell_i) = p_i$. This proves that, in general,

$$\bar{u}_{i1} \neq p_i.$$

We next establish the local expansion. Let

$$f(\eta) := \sigma\big((1+\eta)\ell_i\big).$$

Since σ is smooth, f is three times continuously differentiable in a neighborhood of $\eta = 0$. By Taylor's theorem expanded at $\eta = 0$,

$$f(\eta) = f(0) + f'(0)\eta + \frac{1}{2}f''(0)\eta^2 + R_3(\eta), \tag{29}$$

where the remainder satisfies

$$|R_3(\eta)| \leq C_i|\eta|^3 \tag{30}$$

for some constant $C_i > 0$ depending only on i, provided $|\eta|$ is sufficiently small.

Since

$$f(0) = \sigma(\ell_i) = p_i,$$

and

$$f'(\eta) = \sigma'\big((1+\eta)\ell_i\big)\,\ell_i, \qquad f''(\eta) = \sigma''\big((1+\eta)\ell_i\big)\,\ell_i^2,$$

we obtain at $\eta = 0$*,*

$$f'(0) = \sigma'(\ell_i)\ell_i, \qquad f''(0) = \sigma''(\ell_i)\ell_i^2.$$

Therefore,

$$u_{i1}^{(m)} = p_i + \sigma'(\ell_i)\ell_i\,\eta_{im} + \frac{1}{2}\sigma''(\ell_i)\ell_i^2\,\eta_{im}^2 + R_3(\eta_{im}). \tag{31}$$

Averaging this identity over m *yields*

$$\bar{u}_{i1} = p_i + \sigma'(\ell_i)\ell_i\left(\frac{1}{M}\sum_{m=1}^{M}\eta_{im}\right) + \frac{1}{2}\sigma''(\ell_i)\ell_i^2\left(\frac{1}{M}\sum_{m=1}^{M}\eta_{im}^2\right) + \frac{1}{M}\sum_{m=1}^{M}R_3(\eta_{im}).$$

Using the notation

$$\bar{\eta}_i := \frac{1}{M}\sum_{m=1}^{M}\eta_{im}, \qquad \overline{\eta_i^2} := \frac{1}{M}\sum_{m=1}^{M}\eta_{im}^2,$$

we obtain

$$\bar{u}_{i1} - p_i = \sigma'(\ell_i)\ell_i\,\bar{\eta}_i + \frac{1}{2}\sigma''(\ell_i)\ell_i^2\,\overline{\eta_i^2} + \frac{1}{M}\sum_{m=1}^{M}R_3(\eta_{im}). \tag{32}$$

Finally, since $|\eta_{im}| \le \eta_{\max}$ *for all* m*,*

$$\left|\frac{1}{M}\sum_{m=1}^{M}R_3(\eta_{im})\right| \le \frac{1}{M}\sum_{m=1}^{M}|R_3(\eta_{im})| \le C_i\eta_{\max}^3, \tag{33}$$

which implies

$$\bar{u}_{i1} - p_i = \sigma'(\ell_i)\ell_i\,\bar{\eta}_i + \frac{1}{2}\sigma''(\ell_i)\ell_i^2\,\overline{\eta_i^2} + O(\eta_{\max}^3). \tag{34}$$

This completes the proof.

11 Proof of Theorem 2.

Proof 7 *By Assumption 2,*

$$\mathbf{u}_i^{(m)} = R_{1/\alpha_{im}^\star}(\boldsymbol{\pi}_i). \tag{35}$$

Applying $R_{\alpha_{im}^\star}$ *to both sides and using the Proposition 5 yields*

$$R_{\alpha_{im}^\star}(\mathbf{u}_i^{(m)}) = R_{\alpha_{im}^\star}\big(R_{1/\alpha_{im}^\star}(\boldsymbol{\pi}_i)\big) = R_1(\boldsymbol{\pi}_i) = \boldsymbol{\pi}_i. \tag{36}$$

Therefore, all rectified oracle outputs coincide at $\boldsymbol{\pi}_i$*. According to the closed form of the Hellinger barycenter, their consensus is exactly* $\boldsymbol{\pi}_i$*.*

12 Proof of Lemma 1.

Proof 8 *Fix* $\mathbf{u} \in S_c$ *and* $\alpha \in [a^-, a^+]$*. For each component,*

$$(R_\alpha(\mathbf{u}))_k = \frac{u_k^\alpha}{\sum_{j=1}^K u_j^\alpha} = \frac{e^{\alpha\log u_k}}{\sum_{j=1}^K e^{\alpha\log u_j}}. \tag{37}$$

Differentiating with respect to α *yields*

$$\partial_\alpha(R_\alpha(\mathbf{u}))_k = (R_\alpha(\mathbf{u}))_k\left(\log u_k - \sum_{j=1}^K R_\alpha(\mathbf{u})_j\log u_j\right). \tag{38}$$

Since $\mathbf{u} \in S_c$, we have $u_k \in [c, 1]$ for all k, and therefore

$$|\log u_k| \le |\log c|. \tag{39}$$

It follows that

$$\left| \log u_k - \sum_{j=1}^{K} (R_\alpha(\mathbf{u}))_j \log u_j \right| \le |\log u_k| + \sum_{j=1}^{K} (R_\alpha(\mathbf{u}))_j |\log u_j| \le 2|\log c|, \tag{40}$$

so we can get

$$|\partial_\alpha (R_\alpha(\mathbf{u}))_k| \le 2|\log c|, \qquad \forall k. \tag{41}$$

For fixed $\mathbf{u}$ and k, the map $\alpha \mapsto (R_\alpha(\mathbf{u}))_k$ is continuously differentiable on $[a^-, a^+]$. Hence, by the mean value theorem [1], for any $\alpha, \beta \in [a^-, a^+]$, there exists some λ between α and β such that

$$R_\alpha(\mathbf{u})_k - R_\beta(u)_k = \partial_\gamma R_\gamma(u)_k \big|_{\gamma=\lambda} (\alpha - \beta)$$

and

$$|(R_\alpha(\mathbf{u}))_k - (R_\beta(u))_k| \le \sup_{\gamma \in [a^-, a^+]} |\partial_\gamma R_\gamma(u)_k| \; |\alpha - \beta| \le 2|\log c| \, |\alpha - \beta|.$$

Summing over components gives

$$\|R_\alpha(\mathbf{u}) - R_\beta(\mathbf{u})\|_2 \le 2\sqrt{K} \, |\log c| \, |\alpha - \beta|. \tag{42}$$

Then, we convert this Euclidean bound into a Hellinger bound. For any $\gamma \in [a^-, a^+]$ and any k,

$$(R_\gamma(\mathbf{u}))_k = \frac{u_k^\gamma}{\sum_{j=1}^K u_j^\gamma} \ge \frac{c^{a^+}}{K}, \tag{43}$$

since $u_k \ge c$ and $\sum_{j=1}^K u_j^\gamma \le K$. Therefore both $R_\alpha(\mathbf{u})$ and $R_\beta(\mathbf{u})$ lie in the set $S_{c'}$ with $c' := \frac{c^{a^+}}{K}$. On $[c', 1]$, the derivative of $x \mapsto \sqrt{x}$ is bounded by

$$\left| \frac{d}{dx} \sqrt{x} \right| = \frac{1}{2\sqrt{x}} \le \frac{1}{2\sqrt{c'}}.$$

Applying the mean value theorem componentwise,

$$|\sqrt{p_k} - \sqrt{q_k}| \le \frac{1}{2\sqrt{c'}} \, |p_k - q_k|, \qquad \forall \mathbf{p}, \mathbf{q} \in S_{c'}. \tag{44}$$

Therefore,

$$\|\sqrt{\mathbf{p}} - \sqrt{\mathbf{q}}\|_2 \le \frac{1}{2\sqrt{c'}} \, \|\mathbf{p} - \mathbf{q}\|_2. \tag{45}$$

Using the definition of Hellinger distance,

$$H(\mathbf{p}, \mathbf{q}) = \frac{1}{\sqrt{2}} \|\sqrt{\mathbf{p}} - \sqrt{\mathbf{q}}\|_2 \le \frac{1}{\sqrt{8c'}} \, \|\mathbf{p} - \mathbf{q}\|_2.$$

Applying this to $\mathbf{p} = R_\alpha(\mathbf{u})$ and $\mathbf{q} = R_\beta(\mathbf{u})$ yields

$$H\big(R_\alpha(\mathbf{u}), R_\beta(\mathbf{u})\big) \le \frac{1}{\sqrt{8c'}} \, \|R_\alpha(\mathbf{u}) - R_\beta(\mathbf{u})\|_2 \le \frac{2\sqrt{K} \, |\log c|}{\sqrt{8c'}} \, |\alpha - \beta|. \tag{46}$$

Thus, the claim holds with

$$L = \frac{2\sqrt{K} \, |\log c|}{\sqrt{8c'}} = \sqrt{\frac{K}{2c'}} \, |\log c|, \qquad c' = \frac{c^{a^+}}{K}. \tag{47}$$

13 Proof of Lemma 2.

Proof 9 *Fix $\alpha \in [a^-, a^+]$. For $\mathbf{u} \in S_c$, we have*

$$(R_\alpha(\mathbf{u}))_k = \frac{u_k^\alpha}{\sum_{j=1}^K u_j^\alpha}. \tag{48}$$

Then, we first show that the Jacobian [2] of the map $u \mapsto R_\alpha(\mathbf{u})$ is uniformly bounded on S_c. Let

$$S(\mathbf{u}) := \sum_{j=1}^K u_j^\alpha.$$

Then, we have

$$(R_\alpha(\mathbf{u}))_k = \frac{u_k^\alpha}{S(\mathbf{u})}.$$

Differentiating with respect to $\mathbf{u}_\ell$ and using the quotient rule gives

$$\frac{\partial (R_\alpha(\mathbf{u}))_k}{\partial u_\ell} = \frac{\alpha u_\ell^{\alpha-1} S(\mathbf{u}) - u_k^\alpha \alpha u_\ell^{\alpha-1}}{S(\mathbf{u})^2}. \tag{49}$$

Equivalently,

$$\frac{\partial (R_\alpha(\mathbf{u}))_k}{\partial u_\ell} = \alpha \, \frac{u_\ell^{\alpha-1} \sum_{j=1}^K u_j^\alpha - u_k^\alpha u_\ell^{\alpha-1}}{\left(\sum_{j=1}^K u_j^\alpha\right)^2}. \tag{50}$$

Since $\mathbf{u} \in S_c$, we have $c \le u_k, u_\ell \le 1$. Hence, for every $\alpha \in [a^-, a^+]$,

$$u_k^\alpha \le 1.$$

Moreover, because $\alpha - 1 \in [a^- - 1, a^+ - 1]$ and $x \mapsto x^t$ is monotone in t for each fixed $x \in (0, 1]$, it follows that

$$u_k^{\alpha-1} \le \max\{c^{a^- - 1},\, c^{a^+ - 1}\}, \qquad u_\ell^{\alpha-1} \le \max\{c^{a^- - 1},\, c^{a^+ - 1}\},$$

so

$$S(\mathbf{u}) = \sum_{j=1}^K u_j^\alpha \ge K c^{a^+},$$

Because $u_j \ge c$ and $\alpha \le a^+$,

$$S(\mathbf{u})^2 \ge K^2 c^{2a^+}.$$

It follows that there exists a constant $C = C(c, a^-, a^+) < \infty$ such that

$$\left| \frac{\partial (R_\alpha(\mathbf{u}))_k}{\partial u_\ell} \right| \le C, \qquad \forall\, \mathbf{u} \in S_c,\ \forall\, \alpha \in [a^-, a^+]. \tag{51}$$

Hence the Jacobian of the map $\mathbf{u} \mapsto R_\alpha(\mathbf{u})$ has uniformly bounded operator norm on S_c. Therefore, there exists a constant, still denoted by C, such that

$$\|R_\alpha(\mathbf{u}) - R_\alpha(\mathbf{v})\|_2 \le C \|\mathbf{u} - \mathbf{v}\|_2, \qquad \forall \mathbf{u}, \mathbf{v} \in S_c.$$

Then, we convert this Euclidean Lipschitz bound into a Hellinger bound. Since $\mathbf{u}, \mathbf{v} \in S_c$, we also have

$$u_k \ge c, \qquad v_k \ge c, \qquad \forall k.$$

Furthermore,

$$R_\alpha(\mathbf{u})_k = \frac{u_k^\alpha}{\sum_{j=1}^K u_j^\alpha} \ge \frac{c^{a^+}}{K}, \qquad R_\alpha(\mathbf{v})_k \ge \frac{c^{a^+}}{K},$$

because $u_k, v_k \ge c$ and $\sum_j u_j^\alpha, \sum_j v_j^\alpha \le K$. Thus both $R_\alpha(\mathbf{u})$ and $R_\alpha(\mathbf{v})$ lie in $S_{c'}$ with $c' := \frac{c^{a^+}}{K}$.

On $[c', 1]$, the derivative of $x \mapsto \sqrt{x}$ is bounded by

$$\left| \frac{d}{dx} \sqrt{x} \right| = \frac{1}{2\sqrt{x}} \leq \frac{1}{2\sqrt{c'}}. \tag{52}$$

Hence, for any $\mathbf{p}, \mathbf{q} \in S_{c'}$,

$$\|\sqrt{\mathbf{p}} - \sqrt{\mathbf{q}}\|_2 \leq \frac{1}{2\sqrt{c'}} \|\mathbf{p} - \mathbf{q}\|_2.$$

Using the definition of Hellinger distance,

$$H(\mathbf{p}, \mathbf{q}) = \frac{1}{\sqrt{2}} \|\sqrt{\mathbf{p}} - \sqrt{\mathbf{q}}\|_2 \leq \frac{1}{\sqrt{8c'}} \|\mathbf{p} - \mathbf{q}\|_2. \tag{53}$$

Applying this with $\mathbf{p} = R_\alpha(\mathbf{u})$ and $\mathbf{q} = R_\alpha(\mathbf{v})$ gives

$$H\big(R_\alpha(\mathbf{u}), R_\alpha(\mathbf{v})\big) \leq \frac{1}{\sqrt{8c'}} \|R_\alpha(\mathbf{u}) - R_\alpha(\mathbf{v})\|_2 \leq \frac{C}{\sqrt{8c'}} \|\mathbf{u} - \mathbf{v}\|_2. \tag{54}$$

Finally, since $\mathbf{u}, \mathbf{v} \in S_c$, the map $x \mapsto \sqrt{x}$ is also Lipschitz on $[c, 1]$, and therefore the Hellinger distance and Euclidean distance are equivalent on S_c. In particular, there exists a constant $C' = C'(c) > 0$ such that

$$\|\mathbf{u} - \mathbf{v}\|_2 \leq C' H(\mathbf{u}, \mathbf{v}). \tag{55}$$

Combining the above inequalities yields

$$H\big(R_\alpha(\mathbf{u}), R_\alpha(\mathbf{v})\big) \leq \hat{L} H(\mathbf{u}, \mathbf{v}), \tag{56}$$

for some constant $\hat{L} = \hat{L}(c, a^-, a^+) < \infty$, which proves the claim.

14 Proof of Lemma 3.

Proof 10 *Let $x_m = \sqrt{p_m}, y_m = \sqrt{q_m}$, and define*

$$s = \frac{1}{M} \sum_{m=1}^{M} x_m, \qquad t = \frac{1}{M} \sum_{m=1}^{M} y_m.$$

Since $p_m, q_m \in \Delta^{K-1}$, we have

$$\|x_m\|_2 = \|y_m\|_2 = 1, \qquad x_m, y_m \geq 0. \tag{57}$$

By the closed-form expression of the Hellinger barycenter,

$$\sqrt{\bar{p}} = \frac{s}{\|s\|_2}, \qquad \sqrt{\bar{q}} = \frac{t}{\|t\|_2}. \tag{58}$$

We first show that s and t stay uniformly away from the origin. Since each x_m lies in the nonnegative orthant and satisfies $\|x_m\|_2 = 1$, we have

$$\|x_m\|_1 \geq \|x_m\|_2 = 1.$$

Thus

$$\|s\|_1 = \left\| \frac{1}{M} \sum_{m=1}^{M} x_m \right\|_1 = \frac{1}{M} \sum_{m=1}^{M} \|x_m\|_1 \geq 1. \tag{59}$$

Using $\|z\|_1 \leq \sqrt{K}\, \|z\|_2$, it follows that

$$\|s\|_2 \geq \frac{1}{\sqrt{K}}.$$

Similarly, we can derive

$$\|t\|_2 \geq \frac{1}{\sqrt{K}}.$$

Then, we consider the normalization map $N(z) = \frac{z}{\|z\|_2}$. On the set $\{z \in \mathbb{R}_+^K : \|z\|_2 \geq 1/\sqrt{K}\}$, the map N is Lipschitz with constant at most $2\sqrt{K}$, because for any nonzero z, z',

$$\left\| \frac{z}{\|z\|_2} - \frac{z'}{\|z'\|_2} \right\|_2 \leq \frac{2}{\min(\|z\|_2, \|z'\|_2)} \|z - z'\|_2.$$

Applying this with $z = s$ and $z' = t$ yields

$$\left\| \frac{s}{\|s\|_2} - \frac{t}{\|t\|_2} \right\|_2 \leq 2\sqrt{K} \|s - t\|_2. \tag{60}$$

It remains to bound $\|s - t\|_2$. By convexity of the Euclidean norm,

$$\|s - t\|_2 = \left\| \frac{1}{M} \sum_{m=1}^{M} (x_m - y_m) \right\|_2 \leq \frac{1}{M} \sum_{m=1}^{M} \|x_m - y_m\|_2.$$

Since

$$H(p_m, q_m) = \frac{1}{\sqrt{2}} \|\sqrt{p_m} - \sqrt{q_m}\|_2 = \frac{1}{\sqrt{2}} \|x_m - y_m\|_2, \tag{61}$$

we obtain

$$\|s - t\|_2 \leq \frac{\sqrt{2}}{M} \sum_{m=1}^{M} H(p_m, q_m).$$

Combining the above inequalities gives

$$\|\sqrt{\bar{p}} - \sqrt{\bar{q}}\|_2 = \left\| \frac{s}{\|s\|_2} - \frac{t}{\|t\|_2} \right\|_2 \leq 2\sqrt{K} \cdot \frac{\sqrt{2}}{M} \sum_{m=1}^{M} H(p_m, q_m).$$

Finally, using the definition of Hellinger distance again,

$$H(\bar{p}, \bar{q}) = \frac{1}{\sqrt{2}} \|\sqrt{\bar{p}} - \sqrt{\bar{q}}\|_2 \leq 2\sqrt{K} \cdot \frac{1}{M} \sum_{m=1}^{M} H(p_m, q_m), \tag{62}$$

which proves the claim.

15 Proof of Theorem 3.

Proof 11 *Fix an instance i, and define the ideal matched outputs*

$$\mathbf{u}_{i,\star}^{(m)} := R_{1/\tilde{\alpha}_{im}}(\boldsymbol{\pi}_i), \qquad m = 1, \ldots, M.$$

By the δ-interior assumption and the bounded exponent range, there exists a constant

$$c_0 = c_0(\delta, a^-, a^+) > 0$$

such that

$$\mathbf{u}_{i,\star}^{(m)} \in S_{c_0}, \qquad \forall m.$$

Let the actual rectified outputs be

$$\tilde{\mathbf{u}}_i^{(m)} := R_{\hat{\alpha}_{im}}(\mathbf{u}_i^{(m)}), \qquad m = 1, \ldots, M, \tag{63}$$

and let $\mathbf{q}_i$ denote the Hellinger barycenter of the collection $\{\tilde{\mathbf{u}}_i^{(m)}\}_{m=1}^M$.

For each oracle m, by the triangle inequality,

$$\begin{aligned} H\big(\tilde{\mathbf{u}}_i^{(m)}, \boldsymbol{\pi}_i\big) &= H\Big(R_{\hat{\alpha}_{im}}(\mathbf{u}_i^{(m)}), \boldsymbol{\pi}_i\Big) \\ &\leq H\Big(R_{\hat{\alpha}_{im}}(\mathbf{u}_i^{(m)}), R_{\hat{\alpha}_{im}}(\mathbf{u}_{i,\star}^{(m)})\Big) + H\Big(R_{\hat{\alpha}_{im}}(\mathbf{u}_{i,\star}^{(m)}), \boldsymbol{\pi}_i\Big). \end{aligned}$$

For the first term, Lemma 2 gives

$$H\Big(R_{\hat{\alpha}_{im}}(\mathbf{u}_i^{(m)}), R_{\hat{\alpha}_{im}}(\mathbf{u}_{i,\star}^{(m)})\Big) \le \hat{L}\, H\big(\mathbf{u}_i^{(m)}, \mathbf{u}_{i,\star}^{(m)}\big) \le \hat{L}\, \xi_{im}. \tag{64}$$

For the second term, using $\mathbf{u}_{i,\star}^{(m)} = R_{1/\tilde{\alpha}_{im}}(\boldsymbol{\pi}_i)$ *and Theorem 2, we have*

$$R_{\tilde{\alpha}_{im}}(\mathbf{u}_{i,\star}^{(m)}) = \boldsymbol{\pi}_i.$$

Since $\mathbf{u}_{i,\star}^{(m)} \in S_{c_0}$*, Lemma 1 applies and yields*

$$H\Big(R_{\hat{\alpha}_{im}}(\mathbf{u}_{i,\star}^{(m)}), R_{\tilde{\alpha}_{im}}(\mathbf{u}_{i,\star}^{(m)})\Big) \le L\,|\hat{\alpha}_{im} - \tilde{\alpha}_{im}| \le L\,\Delta_{im}.$$

Therefore,

$$H\big(\tilde{\mathbf{u}}_i^{(m)}, \boldsymbol{\pi}_i\big) \le \hat{L}\,\xi_{im} + L\,\Delta_{im}. \tag{65}$$

Then, consider the ideal reference collection consisting of M *copies of the latent consensus* $\boldsymbol{\pi}_i$*, i.e.,* $\{\boldsymbol{\pi}_i, \ldots, \boldsymbol{\pi}_i\}$*. In the exactly matched case, this is precisely the collection obtained after correct rectification. Hence, its Hellinger barycenter is* $\boldsymbol{\pi}_i$*. Applying Lemma 3 to this ideal reference collection and the actual rectified collection* $\{\tilde{\mathbf{u}}_i^{(m)}\}_{m=1}^M$ *gives*

$$H(\mathbf{q}_i, \boldsymbol{\pi}_i) \le 2\sqrt{K} \cdot \frac{1}{M} \sum_{m=1}^{M} H\big(\tilde{\mathbf{u}}_i^{(m)}, \boldsymbol{\pi}_i\big) \tag{66}$$

$$\le 2\sqrt{K}\hat{L} \cdot \frac{1}{M} \sum_{m=1}^{M} \xi_{im} + 2\sqrt{K}L \cdot \frac{1}{M} \sum_{m=1}^{M} \Delta_{im}. \tag{67}$$

Hence, the claim holds with

$$C_1 := 2\sqrt{K}\hat{L}, \qquad C_2 := 2\sqrt{K}L.$$

16 Proof of Proposition 1.

Proof 12 *For any interior simplex vector* $\mathbf{u}$*, define the log-ratio components relative to component* K *by*

$$\rho_k(\mathbf{u}) := \log \frac{u_k}{u_K}, \qquad k = 1, \ldots, K-1.$$

Under scalar radial rectification,

$$R_\alpha(\mathbf{u})_k = \frac{u_k^\alpha}{\sum_{j=1}^K u_j^\alpha},$$

and therefore

$$\rho_k(R_\alpha(\mathbf{u})) = \log \frac{(R_\alpha(\mathbf{u}))_k}{(R_\alpha(\mathbf{u}))_K} = \log \frac{u_k^\alpha}{u_K^\alpha} = \alpha\, \rho_k(\mathbf{u}). \tag{68}$$

Thus, scalar rectification acts as a common scalar multiplication on the entire log-ratio vector.

Now consider a component-anisotropic distortion of the form

$$T(\mathbf{u})_k = \frac{u_k^{\eta_k}}{\sum_{j=1}^K u_j^{\eta_j}},$$

where $\eta_a \neq \eta_b$ *for some* $a \neq b$*. Then*

$$\rho_k(T(\mathbf{u})) = \log \frac{T(\mathbf{u})_k}{T(\mathbf{u})_K} = \eta_k \log u_k - \eta_K \log u_K. \tag{69}$$

Hence, different log-ratio components are transformed with different effective coefficients. If a scalar rectifier R_α *could invert* T *for all interior* $\mathbf{u}$*, then for every* k *we would need*

$$\rho_k(R_\alpha(T(\mathbf{u}))) = \rho_k(\mathbf{u}).$$

But since R_α multiplies all log-ratios by the same scalar α, this would require a single α to simultaneously undo all component-dependent distortions, which is impossible unless the exponents η_k are all equal. Therefore no scalar radial family can invert the full anisotropic distortion class.

For non-order-preserving distortions, the obstruction is even more immediate. Scalar radial rectification always preserves component order, whereas a non-order-preserving distortion reverses the ordering on some open subset of the simplex. Consequently, no scalar rectifier can undo such an order reversal in general.

17 Supplementary explanation of the experiment.

17.1 Boundary and Tail-Sensitive Validation of Geometric Incomparability.

The results in Tables 6 and 7 examine geometric incomparability from two posterior geometric perspectives: boundary uncertainty and tail-mass allocation. For the first two gaps in Table 6, CFCR and ACR consistently achieve lower HellErr and KLErr than ME, GeoM, and HRaw on both high-gap and low-gap subsets. The improvement on the low-gap samples suggests that comparability recovery remains effective when the two classes are close and predictions are more sensitive to boundaries.

Consistent patterns were also observed across different tail quality ranges in Table 6. The direct fusion baseline remains affected by inconsistent probability assignments outside the dominant class, particularly on the high-tail subset. In contrast, CFCR and ACR significantly reduced the recovery error in most cases, with ACR typically performing best, while CFCR was also competitive in some low-tail cases. These results demonstrate that the proposed correction method can correct not only global sharpness mismatch but also boundary blurring and tail-mass inconsistency.

Table 6: Comparison of consensus errors of CFCR and ACR with different fusion methods on top-two-gap subsets in semantically aligned but geometrically incomparable synthetic scenarios.

Datasets	HellErr ↓					KLErr ↓				
	ME	GeoM	HRaw	CFCR	ACR	ME	GeoM	HRaw	CFCR	ACR
D1_high_gap	2.082	3.554	1.531	0.590	**0.136**	0.517	2.661	0.248	0.029	**0.009**
D1_low_gap	2.285	3.193	1.912	0.975	**0.052**	0.229	0.489	0.162	0.083	**0.000**
D2_high_gap	2.399	4.268	1.994	0.850	**0.294**	0.643	4.193	0.468	0.056	**0.016**
D2_low_gap	2.890	3.728	2.419	1.336	**0.170**	0.357	0.654	0.257	0.157	**0.004**
D3_high_gap	3.776	5.273	3.254	1.814	**1.145**	1.736	7.682	1.484	0.239	**0.119**
D3_low_gap	4.459	4.417	3.567	1.521	**1.034**	0.838	0.874	0.544	0.261	**0.095**
D4_high_gap	3.034	4.457	2.615	1.373	**0.642**	1.170	5.706	0.995	0.139	**0.048**
D4_low_gap	3.327	3.432	2.674	1.109	**0.543**	0.470	0.528	0.309	0.152	**0.032**

17.2 Performance under Different Levels of Initial Oracle Disagreement.

To examine whether the proposed methods remain effective under different degrees of oracle inconsistency, we divide samples into low-, medium-, and high-disagreement subsets according to the discrepancy among the original oracle outputs. Larger disagreement indicates stronger geometric heterogeneity before rectification. HellErr and KLErr are used to evaluate how well the recovered consensus approaches the latent distribution.

As shown in Table 8, the errors of direct fusion methods generally increase as the initial oracle disagreement becomes larger, indicating that raw aggregation is highly sensitive to geometric inconsistency among oracles. In contrast, CFCR and ACR substantially reduce HellErr and KLErr across most subsets, especially under medium- and high-disagreement conditions. This demonstrates that comparability recovery is most beneficial when the original oracle outputs are strongly heterogeneous. ACR performs best in most low- and medium-disagreement cases, while CFCR is more competitive in several high-disagreement settings, suggesting that frozen descriptors can be more stable when the initial mismatch becomes severe.

Table 7: Comparison of consensus errors of CFCR and ACR with different fusion methods on tail-mass subsets in semantically aligned but geometrically incomparable synthetic scenarios.

Datasets	HellErr ↓					KLErr ↓				
	ME	GeoM	HRaw	CFCR	ACR	ME	GeoM	HRaw	CFCR	ACR
D1_low_tail	2.273	3.507	1.523	0.162	**0.098**	0.554	1.905	0.246	0.003	**0.002**
D1_high_tail	2.699	3.506	2.423	0.279	**0.119**	0.384	0.731	0.327	0.007	**0.001**
D2_low_tail	2.575	4.229	2.005	0.396	**0.304**	0.731	3.448	0.450	0.012	**0.011**
D2_high_tail	3.315	4.108	2.868	0.534	**0.308**	0.562	0.999	0.443	0.023	**0.006**
D3_low_tail	4.053	5.866	3.505	1.599	**1.596**	1.903	8.013	1.580	**0.142**	0.181
D3_high_tail	5.034	5.227	4.063	1.528	**1.347**	1.171	1.517	0.793	0.153	**0.115**
D4_low_tail	3.232	4.779	2.763	**0.895**	0.899	1.238	5.376	0.997	**0.050**	0.061
D4_high_tail	3.819	4.062	3.101	0.878	**0.760**	0.693	0.925	0.476	0.051	**0.036**

Table 8: Comparison of consensus recovery errors among CFCR, ACR, and other fusion methods under different levels of initial oracle disagreement in semantically aligned but geometrically incomparable synthetic scenarios.

Datasets	HellErr ↓					KLErr ↓				
	ME	GeoM	HRaw	CFCR	ACR	ME	GeoM	HRaw	CFCR	ACR
D1_low_dis	0.266	0.266	0.197	0.201	**0.005**	0.005	0.005	0.003	0.004	**0.000**
D1_mid_dis	1.503	1.751	1.238	0.120	**0.008**	0.104	0.148	0.073	0.002	**0.000**
D1_high_dis	4.082	7.058	3.244	0.241	**0.116**	0.954	4.702	0.536	**0.005**	0.006
D2_low_dis	0.359	0.333	0.268	0.271	**0.023**	0.009	0.008	0.005	0.007	**0.000**
D2_mid_dis	1.924	2.066	1.542	0.221	**0.046**	0.164	0.198	0.108	0.004	**0.000**
D2_high_dis	5.072	8.451	4.285	0.448	**0.280**	1.346	6.663	0.953	**0.017**	0.020
D3_low_dis	0.500	0.369	0.367	0.611	**0.340**	0.018	0.010	0.010	0.030	**0.008**
D3_mid_dis	2.708	2.301	2.069	0.600	**0.541**	0.314	0.239	0.186	0.028	**0.018**
D3_high_dis	7.835	10.654	6.754	**1.565**	1.858	3.289	12.239	2.685	**0.133**	0.181
D4_low_dis	0.391	0.302	0.289	0.367	**0.152**	0.011	0.006	0.006	0.013	**0.002**
D4_mid_dis	2.055	1.828	1.580	0.337	**0.271**	0.183	0.151	0.110	0.010	**0.004**
D4_high_dis	6.150	8.749	5.301	**0.882**	0.963	2.131	8.934	1.737	**0.055**	0.064

17.3 Descriptor ablation experiment.

We conduct descriptor ablation to evaluate the contribution of different geometry cues in CFCR and ACR. HellErr and KLErr measure distribution-level recovery against the known latent consensus, and are used as the primary criteria. NLL, Brier, and ECE are additionally reported as task-level and calibration-related diagnostics. Lower values indicate better performance for all metrics. The specific definitions are as follows:

$$\mathrm{NLL} = -\frac{1}{n}\sum_{i=1}^{n}\log q_{i,y_i}, \qquad \mathrm{ECE} = \sum_{r=1}^{R}\frac{|\mathcal{B}_r|}{n}\left|\mathrm{acc}(\mathcal{B}_r) - \mathrm{conf}(\mathcal{B}_r)\right|, \tag{70}$$

where $\mathcal{B}_r$ is the r-th confidence bin, and $\mathrm{acc}(\mathcal{B}_r)$ and $\mathrm{conf}(\mathcal{B}_r)$ denote its empirical accuracy and average confidence, respectively.

The descriptor ablation results in Tables 9-12 show that the full refined descriptor consistently provides the strongest or near-strongest distribution-level recovery across D1-D4. In particular, Refined_full achieves the lowest HellErr on all four datasets and also obtains the best or competitive KLErr, confirming the benefit of using consensus-dependent geometry cues. Compared with Frozen_full, the refined version usually yields lower recovery errors, indicating that updating descriptors according to the current consensus improves the estimation of oracle-specific comparability.

Removing individual descriptor components generally weakens recovery performance. The degradation is especially clear when tail-related or boundary-related cues are removed, suggesting that

posterior tail mass and top-two uncertainty are important for correcting geometric mismatch. Some ablated variants obtain lower NLL, Brier, or ECE, but these improvements are not always aligned with HellErr and KLErr. This indicates that sharper task-level predictions may occasionally improve label-based metrics while deviating from the latent consensus. Therefore, the ablation primarily supports the necessity of the full descriptor design from the perspective of probability-level comparability recovery.

Table 9: Ablation study of different descriptor components on D1.

Method	HellErr	KLErr	NLL	Brier	ECE
Frozen_full	0.189	0.004	100.381	48.046	61.400
Frozen_no_entropy	0.568	0.027	**99.466**	**47.619**	**61.054**
Frozen_no_disagreement	0.217	0.005	100.346	48.029	61.380
Frozen_no_tail	0.445	0.013	99.552	47.705	61.170
Refined_full	**0.042**	**0.002**	100.817	48.267	61.572
Refined_no_boundary	0.193	0.005	100.416	48.058	61.406
Refined_no_entropy	0.044	**0.002**	100.811	48.264	61.570
Refined_no_tail	0.144	0.004	100.500	48.120	61.470
Refined_no_disagreement	0.043	**0.002**	100.810	48.264	61.569

Table 10: Ablation study of different descriptor components on D2.

Method	HellErr	KLErr	NLL	Brier	ECE
Frozen_full	0.315	0.009	124.857	56.675	69.510
Frozen_no_entropy	0.687	0.042	123.730	56.216	**69.133**
Frozen_no_disagreement	0.378	0.015	124.720	56.602	69.431
Frozen_no_tail	0.726	0.035	**123.348**	**56.167**	69.169
Refined_full	**0.122**	**0.006**	125.553	56.975	69.730
Refined_no_boundary	0.311	0.011	124.958	56.702	69.518
Refined_no_entropy	0.127	**0.006**	125.531	56.967	69.724
Refined_no_tail	0.352	0.011	124.734	56.652	69.505
Refined_no_disagreement	0.149	**0.006**	125.440	56.922	69.689

Table 11: Ablation study of different descriptor components on D3.

Method	HellErr	KLErr	NLL	Brier	ECE
Frozen_full	0.947	**0.066**	207.231	77.869	86.377
Frozen_no_entropy	1.154	0.095	206.096	77.647	86.230
Frozen_no_disagreement	1.066	0.101	206.593	77.568	86.154
Frozen_no_tail	1.682	0.186	**203.425**	**77.256**	86.004
Refined_full	**0.924**	0.070	207.419	77.861	86.357
Refined_no_boundary	1.030	0.090	206.883	77.645	86.214
Refined_no_entropy	0.956	0.076	207.227	77.800	86.319
Refined_no_tail	1.378	0.138	205.160	77.309	**86.003**
Refined_no_disagreement	1.152	0.116	206.223	77.525	86.130

17.4 Before-and-After Geometric Diagnostics on Real Black-Box Outputs.

We report the oracle-level geometric diagnostics of the real black-box outputs before and after correction in Table 13. Unlike the controlled synthetic setting, the real black-box outputs may involve non-radial, model-specific, and noisy mismatch patterns. Therefore, marginal diagnostics such as EntStd, GapStd, and TailStd do not necessarily decrease on every dataset. For example, AID shows increased marginal dispersion after rectification, indicating that lower dispersion in these one-dimensional summaries is not equivalent to better probability-level comparability.

Table 12: Ablation study of different descriptor components on D4.

Method	HellErr	KLErr	NLL	Brier	ECE
Frozen_full	0.520	0.026	186.003	73.537	83.204
Frozen_no_entropy	0.715	0.042	185.035	73.296	83.035
Frozen_no_disagreement	0.610	0.040	185.572	73.353	83.055
Frozen_no_tail	1.002	0.067	**183.715**	**73.044**	**82.889**
Refined_full	**0.462**	**0.024**	186.319	73.615	83.244
Refined_no_boundary	0.590	0.036	185.752	73.380	83.082
Refined_no_entropy	0.484	0.025	186.212	73.578	83.220
Refined_no_tail	0.816	0.053	184.686	73.176	82.956
Refined_no_disagreement	0.599	0.037	185.647	73.404	83.099

Nevertheless, PairH decreases after rectification on all datasets, suggesting that CFCR and ACR reduce the global Hellinger discrepancy among oracle outputs. This metric is more directly aligned with the Hellinger-based recovery objective than the marginal entropy-, gap-, and tail-based statistics. Averaged across datasets, all four diagnostics also decrease after rectification, with ACR achieving the lowest average EntStd, GapStd, TailStd, and PairH. These results provide supporting evidence that pre-fusion rectification improves global output-space comparability on real black-box outputs, while also reflecting the complexity of real-world geometric mismatch.

Table 13: Oracle-level geometric diagnostics before and after rectification on real-world black-box outputs.

Data	EntStd ↓			GapStd ↓			TailStd ↓			PairH ↓		
	Raw	CFCR	ACR	Raw	CFCR	ACR	Raw	CFCR	ACR	Raw	CFCR	ACR
AID	**16.366**	18.576	18.002	**25.653**	27.005	26.699	**17.079**	19.600	19.015	52.682	52.596	**52.373**
Hate	20.598	**15.979**	16.100	25.827	**21.579**	21.705	0.802	0.760	**0.745**	27.809	**24.322**	24.400
CIF	12.787	6.311	**6.217**	17.586	13.277	**12.858**	16.125	**9.659**	9.663	42.708	40.522	**40.453**
Civi	17.346	**12.952**	13.935	11.273	**8.614**	9.071	**0.000**	**0.000**	**0.000**	19.377	**17.482**	17.806
Net	15.993	16.003	**14.627**	29.151	29.276	**27.905**	12.978	13.586	**12.492**	49.037	47.867	**47.542**
Stan	14.508	14.622	**14.202**	24.213	24.437	**23.792**	**16.732**	17.157	16.785	45.958	**45.905**	45.912
Avg	16.266	14.074	**13.847**	22.284	20.698	**20.338**	10.619	10.127	**9.783**	39.595	38.116	**38.081**

References

[1] Carl Ludwig Siegel. A mean value theorem in geometry of numbers. *Annals of Mathematics*, 46(2):340–347, 1945.

[2] Bogdan M Wilamowski, Nicholas J Cotton, Okyay Kaynak, and GÜnhan Dundar. Computing gradient vector and jacobian matrix in arbitrarily connected neural networks. *IEEE Transactions on Industrial Electronics*, 55(10):3784–3790, 2008.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed this in the article.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the proofs of all theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the detailed experimental settings and the codes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the codes.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the detailed experimental setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have already provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the detailed experimental setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We do not find any societal impact of this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer:[N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used assets are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the proposed method. They are used only as black-box oracles in one application experiment to evaluate the generality of the proposed soft-fusion framework.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.