WHAT SHOULD CONSISTENCY PRESERVE? READOUT-GEOMETRIC INFORMATION TRANSPORT FOR SEMI-SUPERVISED REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Consistency regularization is deceptively underspecified for scalar regression: matching two views does not determine which task information should remain invariant. We show that this missing choice separates useful unlabeled supervision from agreement that distorts the target coordinate or erases label-predictive geometry. We introduce Readout-Geometry Consistency (RGC), a unified framework that factorizes consistency into a readout-induced *transport space* and a *source timescale*. A task can therefore transport either its scalar coordinate or protected within-fiber geometry, using an online or temporally persistent source, through the same objective. We derive when coordinate transport reduces target risk, characterize the variance–lag trade-off of temporal sources, and establish first-order prediction invariance for protected tangent updates. Across image age estimation, text rating prediction, and speech quality assessment, RGC establishes new best reported means on two benchmarks and matches leading performance on the third, while controlled interventions confirm the predicted roles of view support, source timescale, and transport geometry.

1 INTRODUCTION

Semi-supervised regression (SSR) asks a model to exploit abundant unlabeled inputs when labels are continuous and scarce. Most consistency-based methods begin from an appealing premise: two task-preserving views of one example should yield similar predictions. Yet the familiar loss

$$\|f_\theta(x^s) - \mathrm{sg}(f_\xi(x^w))\|^2 \tag{1}$$

commits to three choices: it preserves an absolute scalar, treats the weak view as the source of that scalar, and assumes that both views retain the same target-bearing event. Classification supplies a probability simplex on which consistency can act. Scalar regression has no analogous canonical object: pseudo-target errors have magnitude and direction, and closer predictions need not be closer to the target. Figure 1 makes this ambiguity concrete: matching two predictions specifies the outer form of consistency, but not whether the update should transport the scalar coordinate or geometry within a readout fiber.

Existing SSR methods address uncertainty, ordinal structure, and pseudo-label relations (Dai et al., 2023b; Sun et al., 2025; Huang et al., 2024; Dai et al., 2023a; Wang et al., 2025; Li et al., 2026; Wang et al., 2026). We identify a complementary organizing principle: the *object* transported by consistency must match the target semantics of the views. Controlled crossovers reveal the consequence. Speech quality benefits from transporting within-fiber geometry once both views share the rated event, whereas text rating requires preservation of the scalar coordinate; imposing the complementary geometry collapses its label-predictive structure. The same discrepancy-minimization template can therefore encode either useful invariance or the wrong task information.

We therefore ask: *what task information should consistency preserve?* Our answer is Readout-Geometry Consistency (RGC). For a scalar predictor $f_\theta(x) = w^\top h_\theta(x) + b$, the readout induces a decomposition between (i) the absolute coordinate along w and (ii) relational directions orthogonal to a labeled task subspace. A second axis specifies whether the source is the online model or a temporally persistent encoder. Thus

$$\text{RGC} = \text{transport space} \ \times \ \text{source timescale}. \tag{2}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

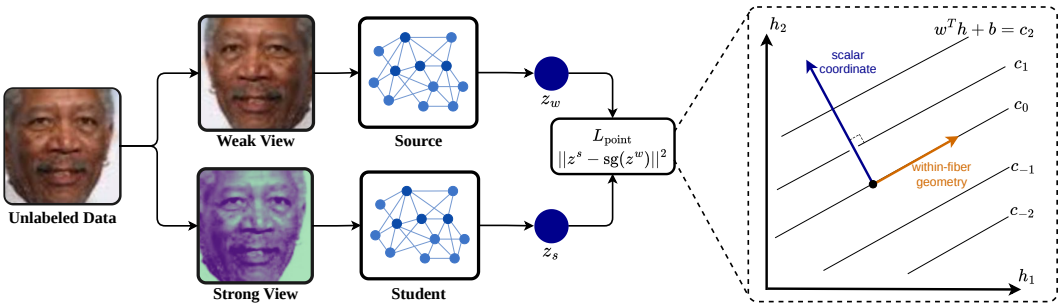


Figure 1: Agreement leaves the transported information object unspecified. The scalar readout separates coordinate from within-fiber geometry; RGC declares which information is preserved.

Each benchmark declares one point in this factorized space through a target-bearing view specification followed by ordinary validation, before any test evaluation. We denote online coordinate, temporal coordinate, and online protected tangent by C, TC, and T, respectively. All three instantiate Eq. 2 and retain the supervised architecture at inference.

In summary, our contributions are as follows:

1. We identify *information-object mismatch* as a central source of unreliable consistency in SSR: agreement is useful only when it preserves task-relevant information.
2. We introduce a unified readout-geometric framework and connect its coordinate, temporal, and protected-tangent instances to target risk through descent, variance–lag, and safety analyses.
3. Across image, text, and audio regression, RGC achieves state-of-the-art or state-of-the-art-level results, with controlled crossovers and representation dynamics validating the predicted transport behavior.

2 Related Work

Consistency learning. Temporal Ensembling and Mean Teacher stabilize predictions over time (Laine & Aila, 2016; Tarvainen & Valpola, 2017), while MixMatch, FixMatch, FlexMatch, FreeMatch, and SoftMatch combine consistency with interpolation, confidence, or adaptive utilization (Berthelot et al., 2019; Sohn et al., 2020; Zhang et al., 2021; Wang et al., 2022b; Chen et al., 2023). UDA further shows that useful consistency depends on semantics of the perturbation itself (Xie et al., 2020; Ghosh & Thiery, 2021). Transferring these ideas to scalar regression is nontrivial because regression has neither a categorical decision boundary nor maximum-class confidence.

Semi-supervised regression. Classical SSR exploits unlabeled manifold geometry (Moscovich et al., 2017); recent deep methods instead model uncertainty, order, or pseudo-label quality. UCVME combines uncertainty consistency with variational ensembling (Dai et al., 2023b); CLSS and GCLSS exploit ordinal structure (Dai et al., 2023a; Wang et al., 2025); RankUp adds an auxiliary ranking classifier (Huang et al., 2024); and HPL learns heteroscedastic pseudo-label weights (Sun et al., 2025). SSR-RCUS preserves local rankings and GaussianMatch filters pseudo-labels through multi-view Gaussian agreement (Li et al., 2026; Wang et al., 2026). They improve pseudo-label quality or relational supervision while typically retaining a fixed notion of what agreement preserves.

Task-relative geometry. Tangent Prop formalized selected invariance (Simard et al., 1991), and manifold regularization connected unlabeled geometry to prediction (Belkin et al., 2006). VICReg addresses representation degeneracy through variance–covariance regularization (Bardes et al., 2021), whereas GradNorm and PCGrad modify gradient scale or interaction (Chen et al., 2018; Yu et al., 2020). These strands motivate controlling invariance, representation geometry, and optimization separately, but do not specify the consistency object induced by a scalar readout.

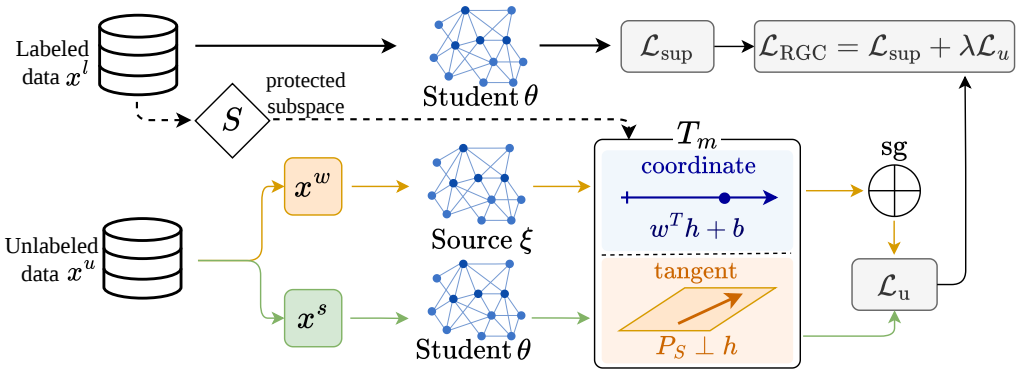


Figure 2: Readout-Geometry Consistency. A task-declared contract specifies support V_m, transport T_m, and source timescale ξ_m within one objective; inference retains the original scalar regressor.

3 Readout-Geometry Consistency

Figure 2 summarizes the common interface. The labeled mini-batch fits the regressor and, for tangent transport, identifies directions that unlabeled consistency must protect. The transport module maps source and student features into the declared readout geometry, detaches the source target, and returns the corresponding consistency loss.

3.1 Problem Setup

Let $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{n_l}$ and $\mathcal{D}_u = \{x_j\}_{j=1}^{n_u}$ be labeled and unlabeled data. A network consists of an encoder $h_\theta : \mathcal{X} \to \mathbb{R}^d$ and a scalar readout

$$f_\theta(x) = w^\top h_\theta(x) + b. \tag{3}$$

For each unlabeled input x, a task-declared view kernel $V_m(\cdot \mid x)$ produces the target-compatible pair (x^w, x^s). A transport contract is the tuple $m = (V_m, A_m, N_m, M_m, \xi_m)$: view support, projection, normalization, metric, and source parameters. Define the transported object and its input-dependent realization as

$$T_m(h) := N_m(A_m h), \qquad z_m(x; \xi_m) := T_m(h_{\xi_m}(x)). \tag{4}$$

Here M_m supplies the quadratic metric used to compare two transported objects; the Euclidean scalar metric and the metric-matched tangent metric are the two instances used below. The unified unlabeled objective is

$$\mathcal{L}_u^{(m)} = \mathbb{E}_{\substack{x \sim \mathcal{D}_u \\ (x^w, x^s) \sim V_m(\cdot|x)}} \left[\| z_m(x^s; \theta) - \text{sg}(z_m(x^w; \xi_m)) \|_{M_m}^2 \right], \tag{5}$$

$$\mathcal{L}_{\text{RGC}} = \mathcal{L}_{\text{sup}} + \lambda_m \mathcal{L}_u^{(m)}. \tag{6}$$

The same supervised loss, model, and evaluation protocol as each benchmark are retained.

3.2 A Readout-Induced Base–Fiber View

The linearized scalar readout partitions representation space into level sets

$$\mathcal{F}_c = \{h \in \mathbb{R}^d : w^\top h + b = c\}. \tag{7}$$

For $w \neq 0$, the representation has the unique orthogonal decomposition

$$h = \frac{ww^\top}{\|w\|_2^2} h + P_{w^\perp} h, \qquad \left[w^\top (h + \delta h) + b\right] - \left[w^\top h + b\right] = w^\top \delta h. \tag{8}$$

Thus $\text{span}(w)$ carries the scalar base coordinate while $\ker(w^\top)$ is exactly the tangent space of each readout fiber. Movement normal to these sets changes the predicted coordinate; movement tangent to

them changes representation geometry without immediately changing the scalar readout. This yields a common geometric language for two forms of consistency. Coordinate transport preserves the *base variable* c, whereas tangent transport preserves relations inside a readout fiber after quotienting directions that labeled data identifies as task-predictive. Source timescale is a second specification axis: either object can be supplied by online or persistent evidence. The axes are separable in definition but need not be additive in effect, because source persistence changes the dynamics under which the transported object is learned.

The factorization does not imply summing all transport losses. Such a sum would assume that both the base coordinate and the fiber geometry are identifiable from every task's views. Instead, RGC instantiates the object supported by the view semantics, an assumption tested by the crossover experiments in Section 5.

3.3 Transport space

The first axis of RGC specifies *what* crosses the weak–strong pair. Both choices are induced by the same scalar readout.

Scalar coordinate. When each view admits the benchmark target, the transported object is the scalar coordinate itself. For a source encoder ξ, we use

$$\mathcal{L}_{\mathrm{coord}}(\xi) = \mathbb{E}\left[\left(f_\theta(x^s) - \operatorname{sg} f_\xi(x^w)\right)^2\right]. \tag{9}$$

Here f_ξ uses the supervised readout with source features. The loss uses every unlabeled example and requires no confidence threshold, sample weight, or auxiliary head.

Relationship to classical consistency. The coordinate instances expose familiar consistency methods as transparent special cases. With an online source, Eq. 9 is ordinary weak–strong output consistency (C). Replacing only the source encoder by its EMA gives encoder-temporal coordinate consistency (TC), a Mean-Teacher-like instance. Their simple form is intentional: the contribution of RGC is not another weighting mechanism, but a minimal specification of the information object and source that the consistency update is permitted to transport.

Label-protected quotient tangent. When absolute pseudo-coordinates are weak but shared-support relations are informative, RGC instead transports tangent geometry outside a labeled task subspace. For a released labeled batch, let $H_l \in \mathbb{R}^{n_l \times d}$ be centered features and $y_l - \bar{y}_l$ centered targets. We construct

$$\mathcal{S} = \operatorname{span}\left\{w,\ \mathbf{P}_{w^\perp} H_l^\top (y_l - \bar{y}_l)\right\}, \tag{10}$$

and let $\mathbf{P}_{\mathcal{S}^\perp}$ denote its orthogonal-complement projector. For view $v \in \{w, s\}$, define

$$\bar{h}_v = \mathbf{P}_{\mathcal{S}^\perp} h_\theta(x^v), \qquad r_v = \max\{\|\bar{h}_v\|_2, \epsilon_{64}\}, \qquad u_v = \frac{\bar{h}_v}{r_v}. \tag{11}$$

The basis of $\mathcal{S}$ is stop-gradient. With B weak features, let $\mu_k = B^{-1}\sum_i h^w_{ik}$ and define the raw weak-feature scale $\sigma_w^2 = \max\{(Bd)^{-1}\sum_{i,k}(h^w_{ik} - \mu_k)^2, \epsilon_{64}\}$. We use the metric-matched angular loss

$$\mathcal{L}_T = \mathbb{E}_i\left[\operatorname{sg}\left(\frac{2 r_i^s r_i^w}{d\sigma_w^2}\right)\left(1 - \langle u_i^s, \operatorname{sg}(u_i^w)\rangle\right)\right], \tag{12}$$

where ϵ_{64} is machine epsilon in the float64 geometric computation. The basis, source object, projected norm factors, and variance scale are all detached; gradients enter only through the strong-view angular term. A vanishing protected direction is omitted, so the basis has rank zero, one, or two without an unstable normalization. We apply this loss only to exact-shared-support views. Protecting only w yields a rank-one readout-only ablation; Eq. 10 additionally protects the main label residual direction revealed by labeled data.

3.4 Source timescale

The second axis specifies *when* the transported object is read. An online source sets $\xi_t = \theta_t$. A persistent source uses an encoder-only exponential moving average (EMA),

$$\xi_t = \alpha\xi_{t-1} + (1-\alpha)\theta_t. \tag{13}$$

Table 1: Task-declared RGC contracts used in our experiments.

Task	Target-bearing unit	View kernel V_m	Candidates	Declared
UTKFace	image instance	same-image weak/strong	C, TC	C
Yelp	review meaning	original/augmented review	C, TC	TC
BVCC	rated segment	exact six-second support	C, TC, T	T

Temporal coordinate transport uses an encoder-only EMA source. Source persistence changes the age of the evidence, not the information object being transported.

3.5 Task contracts and implementation

Equation 6 is shared across modalities. The complete conceptual family is $\mathcal{M}_{\text{full}} = \{\text{coordinate}, \text{protected tangent}\} \times \{\text{online}, \text{temporal}\}$. We use $\mathcal{M} = \{\text{C}, \text{TC}, \text{T}\}$ as the deployment library; the remaining temporal–tangent cell completes matched factorial diagnostics. These symbols index coordinates of the same transport objective, not separate architectures or independently assembled algorithms. A dataset-level contract is declared in two stages. First, the benchmark label defines the target-bearing unit e_t, and only view kernels that retain the same e_t on both sides are admitted. Together with the operator preconditions, this yields the candidate set

$$\mathcal{A}_t = \{m \in \mathcal{M} : V_m \text{ preserves } e_t \text{ and the operator preconditions hold}\}. \tag{14}$$

The standard validation split then fixes one contract

$$\widehat{m}_t \in \arg\min_{m \in \mathcal{A}_t} \widehat{R}_{\text{val}}(\theta_m), \tag{15}$$

The declared contract is reused across seeds, label budgets, and final test evaluation. Each run follows one training trajectory, with no per-example mixture, learned gate, or deployment-time routing.

Yelp and BVCC use their released validation splits in Eq. 15. Because the released UTKFace protocol has no separate validation split, we predeclare its online source and transfer C unchanged across the 50-, 250-, and 2,000-label settings; the public evaluation split is used only for benchmark-aligned checkpoint selection. Hidden unlabeled targets and per-run evaluation outcomes never alter a declaration.

Unified training step. Each update first computes the supervised regression loss and constructs a target-compatible pair of unlabeled views. The weak path supplies a detached object through the declared source; projection, normalization, and metric then define its transport to the strong path. After the hard warm-up, the transport loss is added to the supervised objective and the student is optimized. Tangent transport recomputes and detaches the protected basis from the current labeled minibatch. Temporal coordinate uses an encoder-only EMA source; all source targets are detached. The appendix gives the complete contract-parameterized procedure.

4 When Does Consistency Preserve Task Information?

4.1 Coordinate risk descent

Let z_s and z_w denote strong- and weak-view scalar predictions, y the hidden target, and $e_s = z_s - y$, $e_w = z_w - y$. Consider the idealized update $z'_s = (1-\eta)z_s + \eta z_w$.

Proposition 1 (coordinate descent condition). For squared target risk,

$$(z'_s - y)^2 - (z_s - y)^2 = -2\eta e_s(e_s - e_w) + \eta^2(e_s - e_w)^2. \tag{16}$$

Consequently, decreasing disagreement is beneficial for sufficiently small η only when $\mathbb{E}[e_s(e_s - e_w)] > 0$ with enough margin to dominate the second-order term.

More precisely, let $A = \mathbb{E}[e_s(e_s - e_w)]$ and $B = \mathbb{E}[(e_s - e_w)^2]$. For $B > 0$ and $\eta > 0$, expected target risk decreases exactly when $A > 0$ and $\eta < 2A/B$.

This separates *teacher advantage* from agreement: reducing $|z_s - z_w|$ is useful only when the induced correction is sufficiently aligned with target error. The distinction predicts coordinate transport on UTKFace and the need for target-bearing support alignment on BVCC.

The identity is deliberately a *function-space* statement. For a shared network, let $J_i = \nabla_\theta z_{s,i}$, $g_R = \mathbb{E}_i[e_{s,i} J_i]$, and $g_C = \mathbb{E}_i[(z_{s,i} - z_{w,i}) J_i]$. An optimizer-preconditioned coordinate step changes population risk to first order according to $-\langle g_R, P_t g_C \rangle$, not according to A alone. Cross-example Jacobians, source drift, and the optimizer state therefore determine whether the ideal scalar correction is realizable by one shared parameter update. We evaluate the output-space quantities at the registered preactivation state in all three modalities. Table 7 reports positive A and a nontrivial interval $2A/B$ for every task, verifying the local correction predicted by Proposition 1. The matched training interventions then test the separate realization question under shared parameters.

4.2 Support validity

Let $u(x)$ denote the latent target-bearing event in an input, such as the speech segment whose perceptual quality is rated. Write Y_w and Y_s for the targets of the events retained by the weak and strong views, and let the weak pseudo-target be $Z_w = Y_w + \epsilon_w$.

Proposition 2 (support requirement). The pseudo-target error for the strong-view event obeys

$$\mathbb{E}(Z_w - Y_s)^2 = \mathbb{E}\epsilon_w^2 + \mathbb{E}\Delta^2 + 2\mathbb{E}[\epsilon_w \Delta], \qquad \Delta = Y_w - Y_s. \tag{17}$$

Thus, when weak-source error and event mismatch are uncorrelated, cross-event pairing adds exactly $\mathbb{E}\Delta^2$ pseudo-target risk. This term remains even for a perfect weak predictor and vanishes under exact-shared support.

On BVCC, replacing independent weak/strong crops with exact-shared support restores stable coordinate learning; protected tangent transport then adds a substantial improvement. The controlled intervention separates support validity from transport-space choice.

4.3 Temporal variance–lag trade-off

Proposition 3 (EMA source variance). If online source error is a stationary signal plus independent zero-mean update noise of variance σ^2, the asymptotic EMA noise variance is

$$\mathrm{Var}(e_{\mathrm{EMA}}) = \frac{1-\alpha}{1+\alpha}\sigma^2. \tag{18}$$

The reduction comes with lag relative to a changing signal. More explicitly, if the latent source coordinate satisfies $|\mu_t - \mu_{t-1}| \le v$, then, after initialization transients, the deterministic EMA lag is bounded by

$$|\bar{\mu}_t - \mu_t| \le \frac{\alpha}{1-\alpha} v. \tag{19}$$

Consequently, a sufficient local condition for the temporal source to have lower mean-squared error than its online counterpart is

$$\frac{1-\alpha}{1+\alpha}\sigma^2 + \left(\frac{\alpha v}{1-\alpha}\right)^2 < \sigma^2. \tag{20}$$

A temporal source is therefore useful when coordinate semantics remain valid and online co-adaptation or variance dominates lag. Source timescale and transport space address distinct terms of the consistency problem.

4.4 Protected tangent safety

Proposition 4 (local protected-gradient invariance). With the basis of $\mathcal{S}$ detached during an unlabeled update, $\mathcal{L}_T$ depends on the strong feature only through $\mathbf{P}_{\mathcal{S}^\perp} h^s$. Hence

$$\nabla_{h^s} \mathcal{L}_T \in \mathcal{S}^\perp, \qquad w^\top \nabla_{h^s} \mathcal{L}_T = 0. \tag{21}$$

The feature-space correction induced directly by the tangent loss therefore does not change the current linear readout to first order. Protecting $\mathbf{P}_{w^\perp} H_l^\top (y_l - \bar{y}_l)$ also removes the direct tangent-loss gradient along the leading labeled residual direction. This property is local to the detached basis and current readout; long-horizon rank, prediction-spread, and label-probe dynamics test whether it persists in the nonlinear training trajectory.

Table 2: UTKFace with 250 and 2,000 labels. Bold and underline denote the best and second-best results, respectively.

	250 labels			2,000 labels		
Method	MAE ↓	R^2 ↑	SRCC ↑	MAE ↓	R^2 ↑	SRCC ↑
Supervised	9.42±0.16	0.540±0.014	0.712±0.010	6.28±0.06	0.794±0.004	0.862±0.001
Π-Model (Laine & Aila, 2016)	9.45±0.30	0.534±0.030	0.706±0.015	6.31±0.10	0.790±0.006	0.860±0.003
Mean Teacher (Tarvainen & Valpola, 2017)	8.85±0.25	0.586±0.020	0.745±0.013	6.29±0.03	0.793±0.004	0.862±0.001
CLSS (Dai et al., 2023a)	9.10±0.15	0.586±0.016	0.737±0.014	6.29±0.01	0.794±0.003	0.862±0.001
UCVME (Dai et al., 2023b)	8.63±0.17	0.626±0.006	0.767±0.007	5.90±0.07	0.821±0.007	0.877±0.002
MixMatch (Berthelot et al., 2019)	7.95±0.15	0.692±0.013	0.832±0.008	6.03±0.07	0.824±0.004	0.883±0.002
RankUp (Huang et al., 2024)	7.06±0.11	0.751±0.011	0.835±0.008	5.61±0.07	0.838±0.003	0.887±0.003
RankUp+RDA (Huang et al., 2024)	6.57±0.18	0.782±0.012	0.856±0.005	5.51±0.07	0.844±0.004	0.890±0.003
SSR-RCUS (Li et al., 2026)	6.53±0.15	0.791±0.008	0.856±0.006	5.48±0.05	0.846±0.003	<u>0.894±0.002</u>
GaussianMatch (Wang et al., 2026)	<u>6.38±0.06</u>	<u>0.794±0.005</u>	<u>0.862±0.007</u>	<u>5.36±0.06</u>	<u>0.854±0.004</u>	0.893±0.002
RGC (ours)	**6.144±0.103**	**0.807±0.009**	**0.865±0.006**	**5.234±0.109**	**0.857±0.005**	**0.896±0.004**

4.5 Why task evidence is necessary

The factorized view makes task evidence explicit: unlabeled agreement alone cannot identify which side of a disagreement is closer to the target.

Proposition 5 (no unlabeled-only coordinate certificate). Let a deterministic consistency rule observe (z_s, z_w) and prescribe a nonzero scalar correction $\delta(z_s, z_w)$, without observing y. There exist two target assignments with exactly the same observed pair for which the correction decreases squared risk in one assignment and increases it in the other.

Thus no nontrivial rule based only on student–source disagreement can certify target-risk descent over arbitrary regression tasks. A useful consistency objective must introduce an assumption external to that disagreement—in RGC, target-bearing support, coordinate semantics, source stationarity, or labeled subspace evidence. Proposition 5 therefore provides the identifiability basis for explicit, test-independent task contracts.

Proposition 6 (teacher advantage does not imply parameter descent). There exists a shared-parameter regression problem in which the weak residual has smaller magnitude than the strong residual for every example, yet $\langle g_R, g_C \rangle < 0$. Consequently, descending the point-consistency gradient increases supervised population risk to first order. Per-example teacher accuracy, Proposition 1's ideal output direction, and optimizer-level training utility are therefore distinct objects.

Taken together, the propositions separate three requirements that prediction agreement alone conflates: semantic support, admissible transport, and source stability. Invalid support changes the target-bearing event, unsuitable transport suppresses task information, and an unstable source introduces transient error. We therefore test each factor through matched interventions while holding the remaining training contract fixed.

5 Experiments

5.1 Benchmarks and protocols

We evaluate on the public RankUp benchmark family (Huang et al., 2024): UTKFace-250 for image age estimation (Zhang et al., 2017), Yelp-250 for text rating regression (Asghar, 2016), and BVCC-250 for speech quality assessment (Cooper & Yamagishi, 2021). Each setting uses 250 labeled examples and its released modality-specific training protocol; all RGC results report three-seed means with sample standard deviations. UTKFace follows the released fixed-250 checkpoint-selection and evaluation convention, whereas Yelp and BVCC use validation-selected checkpoints evaluated once on independent test sets. Contracts and central configurations are fixed before final evaluation. Full optimization, implementation, label-budget, and statistical details are provided in the appendix.

Table 3: Text and audio regression with 250 labels. Bold and underline denote the best and second-best SSR results, respectively. '–' denotes not applicable.

	Yelp Review			BVCC		
Method	MAE ↓	R^2 ↑	SRCC ↑	MAE ↓	R^2 ↑	SRCC ↑
Supervised	0.723±0.023	0.566±0.019	0.769±0.010	0.533±0.006	0.490±0.018	0.741±0.009
Π-Model	0.730±0.024	0.565±0.019	0.769±0.009	0.534±0.008	0.489±0.021	0.740±0.009
Mean Teacher	0.730±0.024	0.565±0.019	0.769±0.009	0.532±0.006	0.492±0.018	0.742±0.008
CLSS	0.721±0.010	0.543±0.011	0.748±0.002	0.499±0.010	0.534±0.027	0.748±0.008
UCVME	0.775±0.006	0.540±0.005	0.763±0.005	0.498±0.003	0.553±0.011	0.774±0.008
MixMatch	0.886±0.004	0.381±0.008	0.660±0.004	0.597±0.017	0.353±0.044	0.626±0.031
RankUp	0.661±0.018	0.645±0.013	0.829±0.002	0.470±0.012	0.588±0.028	0.776±0.010
RankUp+RDA	0.632±0.009	0.651±0.007	0.810±0.005	0.463±0.013	0.598±0.027	0.783±0.011
GCLSS	–	–	–	0.453±0.005	0.612±0.007	0.791±0.014
SSR-RCUS	<u>0.579±0.008</u>	<u>0.714±0.006</u>	<u>0.848±0.003</u>	**0.443±0.011**	**0.631±0.025**	<u>0.799±0.012</u>
GaussianMatch	0.630±0.008	0.655±0.009	0.832±0.004	–	–	–
RGC (ours)	**0.572±0.018**	**0.720±0.014**	**0.849±0.007**	<u>0.450±0.012</u>	<u>0.630±0.014</u>	**0.801±0.006**

Table 4: UTKFace with 50 labels. Bold and underline denote the best and second-best results, respectively.

Method	MAE ↓	R^2 ↑	SRCC ↑
Supervised	14.13±0.56	0.090±0.092	0.371±0.071
Π-Model	13.82±1.02	0.100±0.086	0.387±0.092
Mean Teacher	13.92±0.20	0.127±0.037	0.423±0.023
CLSS	13.61±0.92	0.138±0.101	0.447±0.074
UCVME	13.49±0.95	0.157±0.110	0.412±0.127
MixMatch	11.44±0.45	0.401±0.028	0.674±0.035
RankUp	9.96±0.62	0.514±0.043	0.703±0.019
RankUp+RDA	9.33±0.54	0.552±0.041	0.770±0.009
SSR-RCUS	**7.44±0.43**	**0.727±0.042**	**0.824±0.007**
GaussianMatch	8.90±0.54	0.602±0.059	<u>0.776±0.028</u>
RGC (ours)	<u>8.700±0.540</u>	<u>0.626±0.034</u>	0.770±0.032

Table 5: Contract crossovers (MAE↓); †validation diagnostic.

Task	Contract	MAE↓
UTKFace	online coordinate	6.229
UTKFace	tangent	103.56†
Yelp	online coordinate	0.559
Yelp	temporal coordinate	0.551
Yelp	readout-only tangent	0.791
BVCC	shared coordinate	0.526
BVCC	temporal coordinate	0.544
BVCC	temporal tangent	0.549
BVCC	readout tangent	0.465
BVCC	protected tangent	0.463

5.2 Main results

Tables 2 and 3 show strong cross-modal results. RGC leads all metrics on UTKFace and Yelp; on BVCC it improves GCLSS by 0.003 MAE, attains the best SRCC, and is within 0.007 MAE of the leader. These gains arise from task-appropriate contracts without changing the inference architecture.

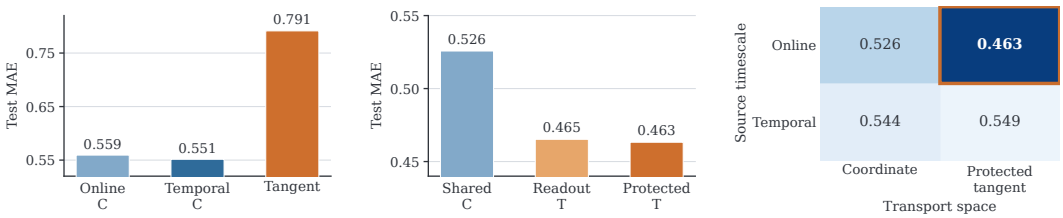
Label efficiency. The same online-coordinate contract leads the 250- and 2,000-label UTKFace settings and remains highly competitive in the 50-label stress test, where it ranks second in MAE and R^2. Across all three budgets, supervision density is the only changed experimental axis; the transport coefficient, warm-up, augmentations, optimizer, update budget, and three-seed protocol are fixed.

5.3 Disentangling support, transport space, and timescale

Table 5 and Figure 3 report *contract crossovers* that change one factor in Eq. 2 while holding the training settings fixed. Coordinate transport suits UTKFace and Yelp, temporal sourcing further stabilizes Yelp, and BVCC benefits from shared event support followed by online protected-tangent transport. The tangent gain reverses under a temporal source, showing that transport space and source timescale form a coupled contract rather than independently additive components.

5.4 Dynamic geometry

To examine how each contract reshapes task information, Figure 4 projects validation features at the common warm-up boundary and endpoint onto checkpoint-specific readout and labeled-residual directions estimated from labeled training data. The UTKFace and BVCC contracts retain broad, target-ordered representations at the endpoint. In contrast, the Yelp readout-only (unprotected) tangent counterfactual compresses prediction spread by more than two orders of magnitude and reduces



(a) Yelp: source timescale. (b) BVCC: online transport space. (c) BVCC: contract interaction.

Figure 3: Readout-geometric crossover. Panels vary source timescale, transport space, and their interaction under matched contracts; highlights indicate the deployed RGC contracts.

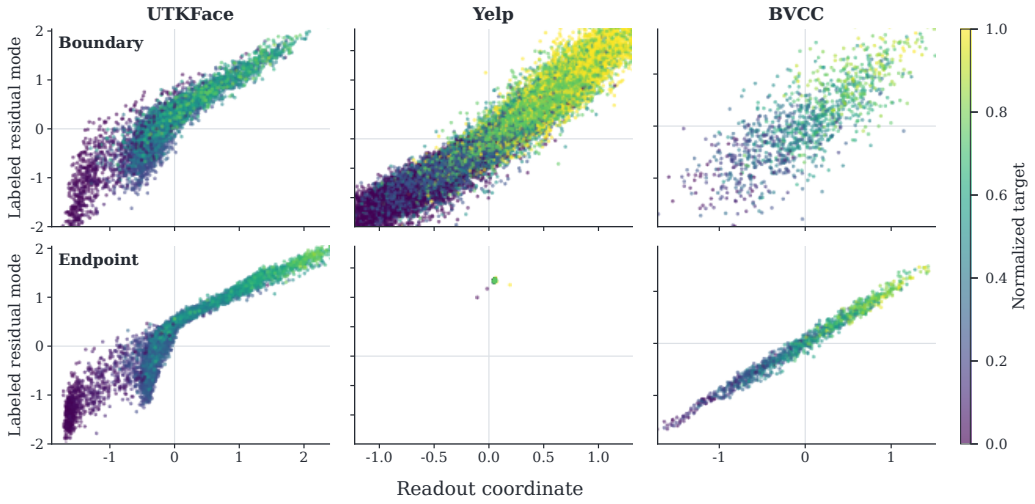


Figure 4: Readout-aligned validation geometry for the registered seed-0 trajectories. Columns denote datasets and rows compare the common warm-up boundary with the fixed endpoint. Color denotes the normalized target and does not enter axis construction.

effective rank from roughly 45 to below 4. The contrast shows that useful consistency preserves label-bearing geometry; agreement can otherwise be achieved through representation collapse.

5.5 Matched controls and paired effects

Matched controls isolate each gain. Exact-shared BVCC crops establish valid event support, and a recipient-matched control rules out parameter routing. With support and seed-matched preactivation states fixed, shared-coordinate transport obtains 0.5073 ± 0.0159 MAE, whereas protected-tangent transport reaches 0.4496 ± 0.0118. The latter improves all three seeds by 0.0577 MAE on average (95% training-seed interval $[0.0474, 0.0680]$). On Yelp, holding coordinate transport fixed and changing only the source timescale improves MAE by 0.0063 (95% paired-example interval $[0.0050, 0.0075]$). Together, these matched interventions validate three contract factors: support validity, the transported information object, and source stability. The resulting dataset-level declaration is reused across runs without per-example routing.

6 Conclusion

Consistency in semi-supervised regression depends on the task information that survives transport. We introduced RGC, a unified objective that specifies this information through a readout-induced transport space and source timescale. Across image, text, and audio regression, RGC achieves SOTA or SOTA-level accuracy, while matched crossovers and representation dynamics verify the roles of event support, source persistence, and protected geometry. These results establish readout geometry as a principled and practical foundation for cross-modal SSR.

Reproducibility Statement

We will release the unified RGC implementation, configurations and evaluation scripts. The appendix specifies every transport contract, augmentation, warm-up, selection protocol, and optimization setting used to reproduce the reported results.

Ethics Statement

We use existing public benchmarks and collect no new human-subject data. UTKFace and BVCC nevertheless contain human faces or voices, and inherited labels may be noisy, demographically imbalanced, or privacy-sensitive. Our claims concern semi-supervised learning behavior under the released research protocols, not biological age determination or deployment-grade speech assessment. We document the known UTKFace duplicate/label-conflict property and do not redistribute raw benchmark data.

AI Use Statement

Generative AI tools were used to assist with language editing and polishing of the manuscript, literature search and review, theory check, debugging of research code, and analysis and interpretation of experimental results. All AI-assisted suggestions, analyses, and outputs were critically reviewed and verified by the authors. The authors determined the final research decisions, methodological choices, experimental conclusions, and manuscript content. The authors take full responsibility for the accuracy, validity and integrity of the work.

References

Nabiha Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7 (11), 2006.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Erica Cooper and Junichi Yamagishi. How do voices from past speech synthesis challenges compare today? *arXiv preprint arXiv:2105.02373*, 2021.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. *Advances in neural information processing systems*, 33:18613–18624, 2020.

Weihang Dai, Yao Du, Hanru Bai, Kwang-Ting Cheng, and Xiaomeng Li. Semi-supervised contrastive learning for deep regression with ordinal rankings from spectral seriation. *Advances in Neural Information Processing Systems*, 36:57087–57098, 2023a.

Weihang Dai, Xiaomeng Li, and Kwang-Ting Cheng. Semi-supervised deep regression with uncertainty consistency and variational model ensembling via bayesian neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7304–7313, 2023b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Atin Ghosh and Alexandre H Thiery. On data-augmentation and consistency-based semi-supervised learning. *arXiv preprint arXiv:2101.06967*, 2021.

Pin-Yen Huang, Szu-Wei Fu, and Yu Tsao. Rankup: Boosting semi-supervised regression with an auxiliary ranking classifier. *Advances in Neural Information Processing Systems*, 37:107444–107468, 2024.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Ximing Li, Jiaxuan Jiang, Changchun Li, You Lu, and Renchu Guan. Semi-supervised regression by preserving ranking relationships between close unlabeled samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23195–23203, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Amit Moscovich, Ariel Jaffe, and Nadler Boaz. Minimax-optimal semi-supervised regression on unknown manifolds. In *Artificial Intelligence and Statistics*, pp. 933–942. PMLR, 2017.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop-a formalism for specifying selected invariances in an adaptive network. *Advances in neural information processing systems*, 4, 1991.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Xueqing Sun, Renzhen Wang, Quanziang Wang, Yichen WU, Xixi Jia, and Deyu Meng. Semi-supervised regression with heteroscedastic pseudo-labels. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 33814–33840. Curran Associates, Inc., 2025. doi: 10.52202/085713-1136.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Ce Wang, Weihang Dai, Hanru Bai, and Xiaomeng Li. Contrastive learning for semi-supervised deep regression with generalized ordinal rankings from spectral seriation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35:3938–3961, 2022a.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022b.

Yin Wang, Hao Lu, Zixuan Wang, Zhen Qin, Li Kuang, Mengchu Zhou, and Shuiguang Deng. Gaussianmatch: Semi-supervised regression with pseudo-label filtering via multi-view gaussian consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31985–31994, 2026.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4352–4360. IEEE, 2017.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Substituting $z'_s = z_s - \eta(z_s - z_w)$ gives $e'_s = e_s - \eta(e_s - e_w)$. Expanding $(e'_s)^2 - e_s^2$ yields Eq. 16. Taking expectations gives the stated condition.

A.2 PROOF OF PROPOSITION 2

Since $Z_w - Y_s = (Z_w - Y_w) + (Y_w - Y_s) = \epsilon_w + \Delta$, squaring and taking expectations gives Eq. 17. If $\mathbb{E}[\epsilon_w \Delta] = 0$, the excess over a same-event weak pseudo-target is $\mathbb{E}\Delta^2$. Exact-shared support gives $Y_w = Y_s$ and therefore $\Delta = 0$ by construction.

A.3 PROOF OF PROPOSITION 3

For independent noise ϵ_t with variance σ^2, the stationary EMA noise is $(1-\alpha)\sum_{k=0}^{\infty} \alpha^k \epsilon_{t-k}$. Its variance is $(1-\alpha)^2\sigma^2 \sum_{k=0}^{\infty} \alpha^{2k} = (1-\alpha)\sigma^2/(1+\alpha)$. For the deterministic component, $|\mu_{t-k} - \mu_t| \le kv$ and hence $(1-\alpha)\sum_{k\ge 0} \alpha^k |\mu_{t-k} - \mu_t| \le (1-\alpha)v \sum_{k\ge 0} k\alpha^k = \alpha v/(1-\alpha)$. Adding the variance and squared worst-case lag gives the sufficient condition in Eq. 20.

A.4 PROOF OF PROPOSITION 4

Let $P = \mathbf{P}_{\mathcal{S}^\perp}$ and write the per-example tangent loss as $\ell(Ph^s)$, with P detached. The chain rule gives $\nabla_{h^s}\ell = P^\top \nabla_{Ph^s}\ell = P\nabla_{Ph^s}\ell$, because an orthogonal projector is symmetric and idempotent. Thus the gradient lies in $\mathcal{S}^\perp$. Since $w \in \mathcal{S}$, $w^\top \nabla_{h^s}\ell = 0$. The same argument holds for the protected labeled residual direction when it is nonzero.

A.5 PROOF OF PROPOSITION 5

Fix an observed pair for which the rule produces $\delta \neq 0$, and let $z_s^+ = z_s + \delta$. If $y_+ = z_s + \delta$, the correction changes squared risk from δ^2 to zero. If instead $y_- = z_s - \delta$, the same correction, under the same observed (z_s, z_w), changes risk from δ^2 to $4\delta^2$. The unlabeled observations cannot distinguish these two target assignments. Therefore a universally risk-improving nonzero correction needs additional task assumptions or labeled evidence.

A.6 PROOF OF PROPOSITION 6

Consider one scalar shared parameter and two equally weighted examples with strong residuals $(10, 1)$, weak residuals $(9, -0.9)$, and parameter Jacobians $(1, -1)$. The weak source is closer to the target on both examples because $|9| < |10|$ and $|-0.9| < |1|$. Ignoring a common positive factor,

$$g_R = \tfrac{1}{2}(10 \cdot 1 + 1 \cdot (-1)) = 4.5, \qquad g_C = \tfrac{1}{2}((10-9)\cdot 1 + (1+0.9)\cdot(-1)) = -0.45. \quad (22)$$

Hence $\langle g_R, g_C \rangle = -2.025 < 0$, so the first-order risk change under $-g_C$ is positive. The counterexample uses neither an inaccurate teacher nor a sign error; the reversal is caused solely by shared-parameter interference.

B THEORY-TO-EVIDENCE CORRESPONDENCE

The propositions in Section 4 make claims at three distinct levels: target risk in output space, admissible directions in representation space, and realization through shared parameters. We therefore test them with matched interventions rather than treating final benchmark accuracy as a proof of mechanism. Table 6 records the prediction attached to each result and the empirical observation that evaluates it.

Matched intervention logic. The crossover experiments change one contract coordinate at a time. The BVCC support intervention changes whether the two views contain the same rated event;

Table 6: Tests and evidence for the predictions induced by each proposition.

Statement	Prediction	Controlled evidence	Supported scope
Coordinate descent	Agreement helps only with sufficient source advantage on the same target-bearing event.	The released-label preactivation probe measures $A > 0$ and $2A/B > 1.8$ in all three tasks; matched trajectories then distinguish coordinate from tangent realization.	Output-space descent and shared-parameter realization are separately testable.
Support requirement	Cross-event views add an irreducible event-mismatch term; exact sharing removes it.	Replacing independent BVCC crops by one copied six-second support restores stable coordinate learning.	Valid support and transport-space choice make separate contributions.
EMA variance–lag	A persistent source can help when the scalar coordinate is meaningful but the online source is noisy or co-adaptive.	On Yelp, only the source timescale changes between online coordinate and encoder-temporal coordinate transport.	Temporal persistence is a coordinate-source option, not a universally superior teacher.
Protected tangent	A detached tangent update is locally orthogonal to the readout and protected labeled residual mode.	In the seed-matched BVCC C–T intervention, protected tangent improves MAE in all three seeds by 0.0577 on average; a stricter readout-only versus label-protected tangent comparison further improves all three seeds by 0.0038 on average. Its trajectory retains prediction spread and label-probe recoverability, while the wrong Yelp unprotected (readout-only) tangent counterfactual removes both.	First-order protection specifies the admissible local direction; dynamics test the long-horizon outcome.
Unlabeled-only identifiability	The same observed disagreement can admit opposite target-risk effects under different hidden targets.	Coordinate, temporal, and tangent routes exhibit sign-changing cross-domain outcomes despite all reducing their declared discrepancies.	Task/view assumptions are necessary; explicit task evidence is part of the RGC contract.
Parameter realization	A favorable per-example output correction need not survive aggregation through shared Jacobians and optimizer state.	The exact two-example construction in Proposition 6 reverses the aggregate gradient despite pointwise source advantage.	Function-space and optimizer-level descent are distinct analytical objects.

after support is shared, the C–T comparison changes only the transported object. The Yelp C–TC comparison retains scalar-coordinate transport and changes only the source timescale. Finally, the boundary-to-endpoint diagnostics in Figures 4 and 5 test whether the local protected-gradient property is realized as sustained label-predictive geometry over training. Together, these interventions connect support, transport space, source timescale, and long-horizon representation dynamics without changing the supervised objective or deployed regressor.

Evidence hierarchy. The analytical results specify testable sufficient conditions and invariants; the crossovers isolate the associated causal factor; the representation probes measure the predicted internal state; and the three-seed benchmark results test whether the resulting contract yields repeatable predictive gains. Agreement between these levels is the basis of the paper's mechanism claim: useful consistency transports information supported by both the task views and the readout geometry.

C Real-Training Certificates and Numerical Invariants

Coordinate output-space probe. At each seed-0 common preactivation state, we draw 16 predeclared view pairs on the 250 released labeled-training examples and compute Proposition 1's $A = \mathbb{E}[e_s(e_s - e_w)]$, $B = \mathbb{E}[(e_s - e_w)^2]$, and $2A/B$. Table 7 reports the median over view draws. No validation or test example enters this diagnostic. The positive values show that the weak prediction defines a locally improving output-space correction on all three tasks. Together with the C–T crossover, this directly separates the output certificate from its long-horizon realization through a shared network, as formalized by Proposition 6. Writing $R(\eta)$ for the squared target risk after an output-space step of size η, Proposition 1 gives $R(\eta) = R(0) - 2\eta A + \eta^2 B$. The final column uses the full matching step, $(R(0) - R(1))/R(0) = (2A - B)/R(0)$.

Protected-tangent invariants in full training. The float64 basis construction and loss in Eqs. 10–12 are instrumented in the three complete BVCC runs. Across 360 logged post-warm-up minibatches per seed, the protected basis has rank two in 100% of records, every logged loss and metric is finite, and the largest displayed inner product between a transported direction and the protected basis is 0.0000 (four-decimal log precision). This real-network trace complements the analytical

Table 7: Proposition 1 quantities at the common preactivation state. The last column is the relative squared-risk reduction of the full output-space match.

Task	A	B	$2A/B$	Rel. reduction
UTKFace	265.1721	264.2174	2.010	97.66%
Yelp	0.7768	0.8451	1.835	97.17%
BVCC	0.1337	0.1376	1.953	96.82%

orthogonality result: the label residual direction is active throughout training, rather than appearing only in a synthetic unit test.

To isolate the labeled direction from tangent transport itself, we run a seed-matched component ablation that changes only the protected space from $\mathrm{span}\{w\}$ to $\mathrm{span}\{w, \mathbf{P}_{w^\perp} H_l^\top (y_l - \bar{y}_l)\}$. The common preactivation checkpoint, exact-shared support, optimizer, update budget, metric matching, and validation selection are unchanged. Table 8 shows that label protection lowers test MAE in all three seeds, with an average gain of 0.0038. This finer intervention complements the larger C–T crossover by attributing a repeatable part of the tangent result specifically to the labeled residual mode.

Table 8: Matched BVCC component ablation (test MAE; lower is better). The two rows differ only in whether the labeled residual direction is included in the protected tangent space.

Protected space	Seed 0	Seed 1	Seed 2	Mean $\pm$ std.
Readout only	0.4652	0.4513	0.4436	0.4534 ± 0.0109
Readout + label residual	0.4632	0.4435	0.4422	0.4496 ± 0.0118

D CONTRACT-PARAMETERIZED ALGORITHM

Every experiment instantiates the contract tuple $m = (V_m, A_m, N_m, M_m, \xi_m)$ in Eq. 6. Table 9 lists the method-defining quantities; the source object and protected projector are stop-gradient in every contract.

Table 9: Operator coordinates for the three task-declared RGC contracts. C, TC, and T denote online coordinate, temporal coordinate, and online label-protected tangent transport. Current denotes the online encoder.

Shorthand	V_m	A_m	N_m	M_m	Weak source ξ_m
C	target-compatible	$w^\top$	identity	scalar Euclidean	current
TC	target-compatible	$w^\top$	identity	scalar Euclidean	encoder EMA
T	exact shared support	$\mathbf{P}_{\mathcal{S}^\perp}$	row ℓ_2	metric matched	current

Given a predeclared tuple, one update is: (i) compute the released supervised regression loss; (ii) construct a target-compatible pair of unlabeled views; (iii) build and detach the labeled protected basis when required; (iv) encode the weak source and strong recipient; (v) apply A_m, N_m, and M_m; and (vi) after the hard warm-up, add the resulting transport loss and update the student. When used, a temporal source is updated once after the student step; each training graph contains exactly one contract.

Unified training step. Let t_0 denote the hard warm-up boundary and let $\mathbb{I}_t$ be one when $t \geq t_0$ and zero otherwise. For a labeled mini-batch B_l and an unlabeled mini-batch B_u, every contract executes the following procedure: for TC, the encoder source is initialized from the student, $\xi_0 = \theta_0$, at the registered preactivation state.

1. Compute $\mathcal{L}_{\mathrm{sup}}$ on B_l with the benchmark's released regression objective.
2. Draw the predeclared target-compatible views (x^w, x^s) for B_u. For T, estimate $\mathcal{S}$ from B_l and detach $\mathbf{P}_{\mathcal{S}^\perp}$ before applying it to unlabeled features.

3. Form
$$q^s = N_m(A_m h_\theta(x^s)), \qquad q^w = \mathrm{sg}[N_m(A_m h_{\xi_m}(x^w))]. \tag{23}$$
4. Minimize
$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \mathbb{I}_t \lambda_m \|q^s - q^w\|^2_{M_m} \tag{24}$$
with respect to the student and scalar regressor.
5. If ξ_m is temporal, update the encoder source after the student step as $\xi_{t+1} = \alpha\xi_t + (1-\alpha)\theta_{t+1}$; otherwise the next step uses the updated online encoder.

The weak transported object is detached in all cases. No contract adds a prediction head, confidence threshold, pseudo-label filter, or inference-time module. Consequently, C, TC, and T differ only in the declared tuple in Table 9, while sharing one optimization interface. For T, both the readout and protected basis are detached, so the unlabeled term satisfies $\nabla_{(w,b)}\mathcal{L}_{\mathrm{T}} = 0$; the scalar head is updated only by $\mathcal{L}_{\mathrm{sup}}$.

Exact tangent numerics. The protected basis is constructed in float64 by modified Gram–Schmidt over the readout and labeled-residual candidates. A candidate whose residual norm does not exceed float64 machine epsilon is omitted. Projection precedes every per-example norm; projected norms and the raw weak-feature variance are clamped below by the same epsilon. The weak object, basis, projected norm factors, and variance factor are detached before the angular loss is formed. These rules fully define the rank-deficient and near-zero cases without a data-dependent threshold.

Dataset-level contract declaration. The procedure in Eqs. 14–15 uses only information available during standard benchmark development. The target-bearing unit and permissible view kernel are written from the dataset label definition, and operator preconditions reduce the fixed library to the candidate set. A released validation split declares one contract when available; otherwise the predeclared online-source default is used. The declaration is copied unchanged to all remaining seeds, label budgets, and test evaluations. Thus validation plays the same role as ordinary dataset-level configuration, rather than learning a gate or choosing a contract per run.

Table 10 makes this declaration step explicit. For Yelp and BVCC, the ordering is computed only on the released validation split of the registered development runs. UTKFace is not ranked using its public evaluation split: its online-coordinate contract is the predeclared fallback described in Section 3.

Table 10: Validation-only audit of dataset-level contract declaration. The relations compare validation MAE (lower is better); final test outcomes are never queried by the declaration rule.

Task	Candidates	Validation-only evidence	Declared	Test used?
UTKFace	C, TC	predeclared online fallback	C	No
Yelp	C, TC	TC $<$ C	TC	No
BVCC	C, TC, T	T $<$ C, TC	T	No

E Additional Experimental Details

All implementations extend the USB-derived public RankUp codebase (Wang et al., 2022a). The modality-specific encoders follow the public benchmark: WRN-28-2 for UTKFace (Zagoruyko & Komodakis, 2016), BERT-Small for Yelp (Devlin et al., 2019), and Whisper-base for BVCC (Radford et al., 2023). Yelp and BVCC use AdamW (Loshchilov & Hutter, 2017); UTKFace retains the released SGD contract.

Comparison scope. The main tables use the established cells associated with each benchmark method; causal claims are instead based on the matched continuations in Section 5, which share initialization, support, optimizer, and update budget. GCLSS is directly included for BVCC because it reports the RankUp split, 250 labels, Whisper-base, and 102,400 updates, obtaining 0.453 ± 0.005 MAE and 0.612 ± 0.007 R^2 over six seeds (Wang et al., 2025). Its UTKFace study uses the 21–60 age subset, ResNet-50, label fractions, and a 30-epoch protocol, so those cells are not mixed with our released WRN label-budget table. HPL likewise evaluates a different UTKFace/IMDB-WIKI/STS-B

percentage-label matrix (Sun et al., 2025). As a supplementary same-codebase check, our Yelp-250 seed-0 port of HPL obtains 0.5864 independent-test MAE, compared with 0.5515 for RGC-TC under the same labeled subset and evaluation split. We keep this single-seed adaptation outside the three-seed main table.

Splits and labeled subsets. The released UTKFace protocol contains 18,964 training and 4,741 evaluation images. Yelp contains 250,000/25,000/50,000 train/validation/test reviews, and BVCC contains 4,974/1,066/1,066 train/validation/test utterances. Fixed index files select the labeled subset for each seed; the remaining training protocol, including the full unlabeled pool, is unchanged. The 50-, 250-, and 2,000-label UTKFace experiments use the corresponding seed-matched index files, while Yelp and BVCC use 250 labeled examples. Thus a seed jointly fixes the labeled subset and registered preactivation state, and all matched controls reuse both.

Table 11: Frozen optimization contracts. Evaluation EMA is the benchmark's checkpoint predictor and is distinct from the unlabeled source. Batch is labeled batch $\times(1+\mu)$, where μ is the unlabeled ratio.

Task	Encoder	Updates	Batch	Optimizer (LR / WD)	Eval EMA
UTKFace	WRN-28-2	262,144	$32(1+7)$	SGD ($10^{-2}/10^{-3}$)	0.999
Yelp	BERT L4-H512-A8	102,400	$8(1+1)$	AdamW ($10^{-5}/5\times10^{-4}$)	none
BVCC	Whisper-base	102,400	$8(1+1)$	AdamW ($2\times10^{-6}/2\times10^{-5}$)	none

All tasks use the released L1 supervised objective and activate consistency after a hard 10% warm-up. Coordinate transport uses $\lambda = 15$ on UTKFace and Yelp; the metric-matched tangent loss has its analytically fixed unit coefficient on BVCC. Yelp uses an encoder-only source EMA with $\alpha = 0.999$; inference retains the online scalar regressor. The UTKFace framework EMA is used only as the released evaluation predictor and does not supply the weak-view target.

View contracts. UTKFace uses the released weak random crop/flip and an independently sampled strong crop/flip followed by RandAugment$(3, 5)$ (Cubuk et al., 2020). Yelp pairs the released original review with one of its two preprocessed augmented reviews. BVCC draws one random six-second support and copies the same waveform values into both network forwards; no second crop or destructive waveform transform is applied. This exact-shared construction preserves the utterance event while retaining independent training-mode network stochasticity.

Continuation and selection. For reproducible causal comparisons, every method instance resumes from the registered seed-matched checkpoint immediately preceding its consistency activation and retains the original optimizer, scheduler, RNG, and model state. Before the coefficient becomes nonzero, the framework may still construct unlabeled forwards and update stateful normalization; we therefore refer to this as a *common preactivation state*, not a pure-supervised trajectory. Yelp and BVCC select checkpoints every 1,024 updates on an independent validation set and construct the test loader only for the final frozen-checkpoint evaluation. UTKFace follows the released RankUp fixed-250 checkpoint-selection and evaluation convention without modification, ensuring a protocol-aligned comparison with the published baseline table.

Metrics and aggregation. For targets y_i and predictions $\hat{y}_i$, we report $\mathrm{MAE} = n^{-1}\sum_i |y_i - \hat{y}_i|$ and $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$; SRCC is Spearman's rank correlation between y and $\hat{y}$. Lower MAE and higher R^2/SRCC are better. Main tables report the arithmetic mean and sample standard deviation over three seeds. Each seed is checkpoint-selected under the protocol stated above and evaluated once under its fixed test convention. The per-seed values underlying the principal three-domain result are listed in Table 12.

Representation diagnostics. All geometry diagnostics use released labeled-training examples to fit probes and the validation split only for evaluation; test examples are never used. For centered validation features H with singular values $\{s_j\}$, define $p_j = s_j^2 / \sum_k s_k^2$ and

$$r_{\mathrm{eff}}(H) = \exp\left(- \sum_{j:p_j>0} p_j \log p_j \right). \tag{25}$$

Table 12: Per-seed RGC results underlying the three-domain summaries. Selected update is the checkpoint chosen by the benchmark-aligned evaluation protocol.

Task	Contract	Seed	MAE ↓	R^2 ↑	SRCC ↑	Selected update
UTKFace-250	C	0	6.2290	0.8009	0.8586	73,728
	C	1	6.0290	0.8171	0.8699	63,488
	C	2	6.1735	0.8030	0.8678	123,904
Yelp-250	TC	0	0.5515	0.7354	0.8566	98,304
	TC	1	0.5842	0.7070	0.8432	102,400
	TC	2	0.5815	0.7179	0.8486	92,160
BVCC-250	T	0	0.4632	0.6150	0.7997	38,912
	T	1	0.4435	0.6424	0.8078	37,888
	T	2	0.4422	0.6339	0.7959	27,648

Table 13: Per-seed UTKFace label-budget extensions. All rows use the same C contract and the released evaluation convention; only the labeled subset and its seed-matched preactivation state change.

Labels	Seed	MAE ↓	R^2 ↑	SRCC ↑	Selected update
50	0	8.3610	0.6466	0.7760	99,328
	1	8.4170	0.6439	0.7986	129,024
	2	9.3228	0.5869	0.7347	135,168
2,000	0	5.1100	0.8630	0.9004	206,848
	1	5.3143	0.8526	0.8921	139,264
	2	5.2782	0.8548	0.8940	120,832

The label mode is the normalized readout-complement cross-covariance $v_l \propto \mathbf{P}_{w^\perp} H_l^\top (y_l - \bar{y}_l)$. Its reported R^2 is obtained by fitting a one-dimensional affine probe on the labeled-training projection $H_l v_l$ and evaluating it on validation features from the same checkpoint. Prediction spread is the standard deviation of native validation predictions. "Boundary" in Figures 4 and 5 denotes the registered end of the supervised warm-up, immediately before consistency is activated; "Endpoint" denotes the fixed final update. Probe definitions, data splits, and checkpoint semantics are identical across the two curves.

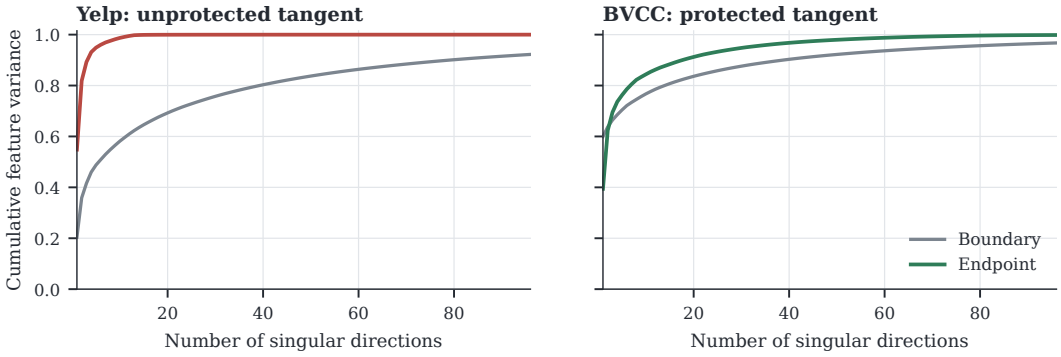


Figure 5: Cumulative validation-feature spectrum from the same boundary and endpoint checkpoints as Figure 4. The Yelp unprotected (readout-only) tangent counterfactual concentrates variance into very few directions, whereas the BVCC protected-tangent endpoint retains a broad spectrum close to its boundary state.

Resource cost. All three contracts retain the supervised regressor at inference. Coordinate and tangent transport add no learned parameters; temporal coordinate transport maintains an encoder-only EMA during training and discards it at deployment. On an RTX 4080 SUPER, the method-specific continuation averaged 6:13, 2:53, and 1:52 hours for UTKFace, Yelp, and BVCC, respectively, across

three seeds. These wall times exclude the reusable 10% common preactivation prefix and are reported per benchmark because backbone, input length, and update budget differ.

F Sensitivity and Paired Statistics

We vary one continuous quantity around each contract while keeping the transport object, data, seed, and full update budget fixed: coordinate coefficient $\{7.5, 15, 30\}$ on UTKFace, encoder-source decay $\{0.99, 0.999, 0.9999\}$ on Yelp, and tangent coefficient $\{0.5, 1, 2\}$ on BVCC. The central settings and main-table checkpoints were fixed before evaluating their local neighborhoods. For mechanism contrasts sharing an ordered evaluation loader, we additionally report seed-stratified paired MAE/MSE bootstrap intervals with 10,000 replicates. Published cross-paper aggregates are never treated as paired.

Table 14: One-seed local sensitivity of validation-selected checkpoints (test MAE; lower is better). UTKFace uses the released evaluation convention; Yelp and BVCC use their independent test sets. The center value is the main configuration.

Task	Scalar	Low	Main	High
UTKFace	λ_C	6.7003	6.2290	6.0884
Yelp	α	0.5528	0.5515	0.5694
BVCC	λ_T	0.4787	0.4632	0.4739

RGC retains strong accuracy throughout these neighborhoods without changing the transport contract. Yelp is stable across the two shorter-memory sources, BVCC is strongest at the central tangent coefficient, and UTKFace improves smoothly with stronger coordinate pressure. The response curves show that each contract behaves as a continuous training objective rather than a brittle discrete choice.

Table 15: Within-codebase paired RGC gains over matched controls. Positive values favor RGC. Yelp intervals resample paired test examples within each seed and weight the three seeds equally; BVCC intervals are two-sided training-seed t intervals over three seed-matched continuations.

Contrast	Metric	Seeds	Mean gain [95% CI]	Positive seeds
Yelp TC vs. online C	MAE	3	0.0063 [0.0050, 0.0075]	3/3
Yelp TC vs. online C	MSE	3	0.0362 [0.0322, 0.0403]	3/3
BVCC T vs. shared-support C	MAE	3	0.0577 [0.0474, 0.0680]	3/3
BVCC T vs. shared-support C	MSE	3	0.0710 [0.0426, 0.0994]	3/3

The paired effects isolate both halves of the RGC principle. On Yelp, temporal-coordinate transport improves MAE in every seed, directly validating the encoder-timescale source. On BVCC, protected tangent transport likewise improves every seed over the exact-shared coordinate control, with an average MAE reduction of 0.0577. Because support, initialization, and training are matched, this persistent gain is attributable to the transported subspace rather than shared-support augmentation alone.