Dual-Granularity Learning for Regression with Continuous Noisy Labels

Hui Guo[1] **Boyu Wang**[1,3] **Grace Y. Yi**[1,2,3,*]

[1]Department of Computer Science, University of Western Ontario
[2]Department of Statistical and Actuarial Sciences, University of Western Ontario
[3]Vector Institute

hguo288@uwo.ca bwang@csd.uwo.ca gyi5@uwo.ca

[*]Corresponding author

Abstract

Learning with noisy labels has been extensively studied for classification, yet its regression counterpart remains relatively underexplored. Unlike classification, regression involves continuous-valued targets where label noise exhibits heterogeneous magnitudes across instances, which poses conflicting requirements on the model: correcting small-to-moderate deviations requires sensitivity to label values for instance-level refinement, whereas large deviations caused by corrupted or outlying annotations render exact label values unreliable, necessitating reduced reliance on exact labels and instead leveraging ordinal structure among samples. These requirements call for methods that capture both global structure and sample-specific deviations in a unified manner. To this end, we propose GRAIN, a dual-granularity framework for robust noisy regression. At the fine-grained level, we derive an explicit residual learning mechanism from a robust loss formulation, introducing instance-specific auxiliary variables to decouple label noise from predictive signals. At the coarse-grained level, we introduce a binning-based contrastive regularization that enforces an order-preserving feature structure aligned with label proximity, addressing noise beyond the capacity of local residual refinement. The two levels are coupled via a histogram-based bridging mechanism that enables bidirectional information flow: coarse-grained structural guidance establishes a well-ordered feature space that anchors residual learning, while fine-grained residual refinement corrects instance-level deviations that structural regularization alone cannot resolve. Experimental results show that GRAIN consistently improves robustness and generalization under diverse noise conditions. Code is available at https://github.com/hguo1728/GRAIN.

1 Introduction

Regression is a fundamental learning paradigm with broad applications in multiple fields, including finance, healthcare, and economics. In practice, however, regression models are often trained with imperfect supervision, where noisy labels arise inevitably due to measurement errors, annotation uncertainty, or system bias [1]. Such label noise can destabilize training, degrade predictive reliability, and substantially limit the deployment of regression systems in high-stakes applications.

Despite its practical importance, learning with continuous noisy labels remains relatively underexplored compared to classification [2, 3, 4]. This gap stems from a fundamental difference in label structure: regression labels are continuous-valued and lack explicit decision boundaries, leading to label noise with heterogeneous magnitude across instances. This induces conflicting requirements

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

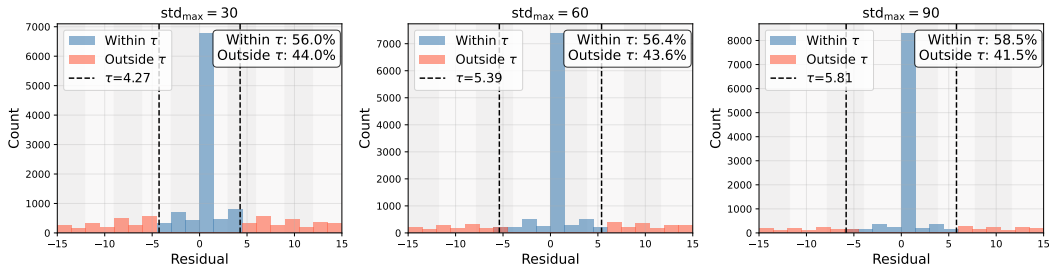


Figure 1: Distribution of true label noise $\{\widetilde{Y}_i - Y_i : i = 1, \ldots, n\}$ at the final training epoch on UTKFace under three noise levels ($\text{std}_{\max}$ = 30, 60, 90). The **dashed lines** (**- - -**) indicate $\pm\tau$. **Blue bars** (■) show noise within $(-\tau, \tau)$ (within the residual head's capacity), and **red bars** (■) show noise outside $(-\tau, \tau)$ (beyond the residual head's capacity). **The proportion of label noise** within and outside $(-\tau, \tau)$ is annotated in each panel. **Alternating shaded regions** represent consecutive label bins. The label noise range is clipped to $[-15, 15]$ for visualization purposes.

on how a model should respond to corrupted supervision. For small-to-moderate deviations, label values are imperfect but retain sufficient information for local correction, requiring *sensitivity to label values* for precise instance-level refinement. For large deviations, which often arise from outlying or heavily corrupted annotations, exact label values become unreliable and may mislead learning, necessitating *reduced reliance on label values* and instead modeling and enforcing the underlying ordered feature-label structure. These conflicting requirements cannot be effectively resolved by a single mechanism. Existing approaches largely focus on improving overall robustness [5, 6, 7] without explicitly addressing these coupled challenges.

To address these challenges, we propose **GRAIN**, a dual-granularity framework that explicitly models instance-level residual noise while enforcing a globally ordered feature-label structure within a unified learning process. At the *fine-grained level*, we derive a piecewise robust loss family for noisy regression and identify a formulation admitting an infimal convolution representation, which introduces auxiliary variables that explicitly capture sample-wise residuals with heterogeneous magnitudes. Based on this formulation, we design a lightweight residual head to learn these variables during training, enabling label noise to be explicitly separated from predictive signals. To prevent the residual head from absorbing predictive signals, we impose an adaptive capacity constraint controlled by a threshold τ, such that it is restricted to modeling only bounded residual components. As illustrated in Figure 1, this constraint naturally induces a separation of correction regimes: deviations within $(-\tau, \tau)$ are handled through fine-grained residual refinement, while larger deviations, where label values become unreliable and induce structural distortion in feature space, are instead addressed by coarse-grained structural regularization. Accordingly, at the *coarse-grained level*, we address the remaining structural corruption by introducing a binning-based contrastive objective. Specifically, we perform adaptive binning based on the density of noisy labels and encourage representations of adjacent bins to align while pushing apart distant bins, thereby preserving the ordinal structure of the label space. This design enforces global order in the feature space while remaining robust to large-magnitude noise that cannot be captured by local residual refinement.

As illustrated in Figure 2, these two mechanisms operate at different granularities and are inherently complementary. Fine-grained refinement provides precise instance-level correction by capturing localized residual variations, while coarse-grained structuring enforces an order-preserving feature geometry aligned with label proximity. A histogram head operating over the binned label space serves as a mediator, extracting pseudo labels from the shared feature representations to provide coupled supervision across the two levels: the pseudo labels guide coarse-grained bin assignment and provide supervision for fine-grained residual learning. In turn, the refined and structured feature representations further improve pseudo-label quality, forming a closed-loop between feature learning and pseudo-label estimation. Together, this unified design provides a coherent solution to noisy regression by simultaneously addressing local corruption and global structural distortion.

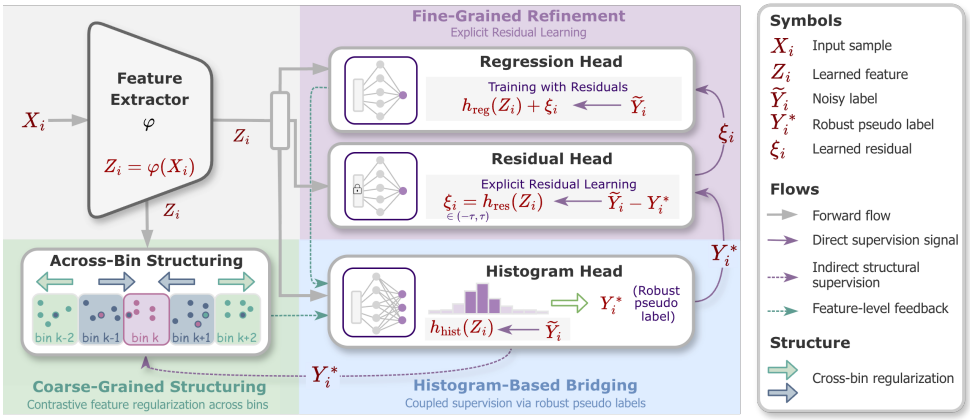


Figure 2: Overview of **GRAIN**. **Fine-grained refinement** (top right): a regression head h_{reg} produces predictions corrected by instance-specific residuals ξ_i learned by a capacity-constrained residual head $h_{\text{res}} \in (-\tau, \tau)$, decoupling sample-wise noise from predictive signals. **Coarse-grained structuring** (bottom left): a binning-based contrastive objective enforces an order-preserving feature structure aligned with label proximity. **Histogram-based bridging** (bottom right): a histogram head h_{hist} generates robust pseudo labels Y_i^*, guiding bin assignment and providing residual supervision targets, while benefiting in turn from the representations refined by both modules.

2 Related Work

Learning with Noisy Labels. Learning with noisy labels has been extensively studied in classification, where labels are discrete and noise is often addressed through sample selection [2, 8], self-training [4], as well as annotation modeling and robust pseudo-labeling [9, 10, 11]. In contrast, learning with continuous noisy labels for regression has received considerably less attention. Existing approaches primarily focus on improving robustness at the loss or sample level, for example via robust loss functions [12], data augmentation [6], adaptive reweighting [7], or sample selection [13]. While effective in reducing sensitivity to noisy labels, these methods treat residuals implicitly as errors rather than deriving explicit residual variables from the loss formulation itself, and typically handle label corruption in a uniform manner, without accounting for its heterogeneous magnitude across instances or its effect on the global structure of the feature space. Unlike these approaches, our method jointly addresses instance-level residual noise and global structural distortion within a unified dual-granularity framework.

Discretization / Binning-based Methods. Discretization-based auxiliary classification has been explored in regression to stabilize training and improve representation learning [14, 15]. These methods map continuous targets into discrete bins and apply classification-style supervision, which can enhance feature separability but often ignores the intrinsic ordering and continuity of regression labels. Instead, our approach leverages adaptive binning as a mechanism for structural regularization, enforcing order-preserving relationships across bins to establish global feature structure under noisy supervision rather than redefining the task as classification. Furthermore, a histogram head operating over the same bin structure generates robust pseudo labels that bridge coarse-grained structural guidance and fine-grained residual refinement. Discretized density guidance has also been studied for source-free regression adaptation [16]; GRAIN instead targets continuous label corruption within a single training distribution.

3 Preliminaries

We consider a regression setting with a p-dimensional input random variable X and an associated label Y, where $\mathrm{X} \in \mathcal{X}$ and $\mathrm{Y} \in \mathcal{Y}$, with $\mathcal{X} \subset \mathbb{R}^p$ denoting the input space and $\mathcal{Y} \subset \mathbb{R}$ the label space. In many applications, the true label Y is not directly observable. Instead, one has access to a noisy label $\widetilde{\mathrm{Y}} \in \mathcal{Y}$. The data are assumed to follow the generative mechanism:

$$\mathrm{Y} = f(\mathrm{X}; \theta^*) + \varepsilon', \tag{1}$$

$$\widetilde{\mathrm{Y}} = \mathrm{Y} + \varsigma(\mathrm{X}, \mathrm{Y})\varepsilon'', \tag{2}$$

where $f(\cdot;\theta^*)$ denotes the unknown ground-truth regression function, indexed by parameter θ^*; ε' and ε'' are independent zero-mean noise variables with finite variance, independent of X; and $\varsigma : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ represents an unknown noise scale function, potentially instance and label-dependent, allowing for heteroskedastic and structured label corruption.

Let $\mathcal{D} = \{\mathrm{X}_i, \widetilde{\mathrm{Y}}_i\}_{i=1}^N$ denote a training sample of size N, where $\widetilde{\mathrm{Y}}_i$ is the noisy label associated with X_i, which may differ from the unobserved true label Y_i for $i \in [N]$. We assume that $(\mathrm{X}_i, \mathrm{Y}_i, \widetilde{\mathrm{Y}}_i)$ are independent and identically distributed (i.i.d.) according to the joint distribution of $(\mathrm{X}, \mathrm{Y}, \widetilde{\mathrm{Y}})$ characterized by (1)-(2). Our goal is to learn a predictor that approximates the underlying regression function $f(\cdot;\theta^*)$ using only noisy data $\mathcal{D}$, which achieves reliable generalization on unseen inputs.

4 GRAIN: A Dual-Granularity Framework for Noisy Regression

We present GRAIN, a dual-granularity framework for deep regression under noisy labels (Figure 2), designed to effectively address the conflicting requirements induced by heterogeneous label noise. An adaptive capacity threshold τ demarcates two regimes: deviations within $(-\tau, \tau)$ are addressed by fine-grained residual refinement, while larger deviations are delegated to coarse-grained structural regularization. Section 4.1 develops a piecewise robust loss family with generalization guarantees, and derives an infimal convolution reformulation that introduces instance-specific auxiliary variables for fine-grained residual modeling within the capacity bound $(-\tau, \tau)$. Section 4.2 introduces a binning-based contrastive regularization that operates at a coarser granularity, enforcing an order-preserving feature geometry across bins of the label space and addressing structural corruption beyond the reach of local refinement. Section 4.3 describes how a histogram-based bridging mechanism mediates supervision across both levels, coupling residual targets and bin assignments through shared pseudo labels to form a closed-loop between feature learning and noise correction.

4.1 Fine-Grained Refinement: Loss-Induced Residual Learning

We begin by focusing on robust prediction given a feature representation, with robust feature learning under noisy supervision deferred to Section 4.2. Let $\mathrm{Z} = \varphi(\mathrm{X})$ denote the feature representation of input X, where $\varphi : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^d$ embeds the input into a d-dimensional feature space $\mathcal{Z}$. We assume that φ is sufficiently expressive such that the regression structure in (1) admits a linear parameterization over $\mathcal{Z}$: $\mathrm{Y} = \langle \mathrm{Z}, \theta^*_{\text{reg}} \rangle + \varepsilon'$, where θ^*_{reg} represents the resulting regression parameter, and $\langle a, b \rangle$ denotes the dot product of vectors a and b. The training dataset is then $\mathcal{D}^\dagger = \{\mathrm{Z}_i, \widetilde{\mathrm{Y}}_i\}_{i=1}^N$, with $\mathrm{Z}_i = \varphi(\mathrm{X}_i)$ for all $i \in [N]$, and let $\widehat{P}^\dagger = \frac{1}{N}\sum_{i=1}^N \delta_{\mathrm{Z}_i, \widetilde{\mathrm{Y}}_i}$ denote the empirical distribution of $(\mathrm{Z}, \widetilde{\mathrm{Y}})$ induced from dataset $\mathcal{D}^\dagger$, where $\delta_{\mathrm{Z}_i, \widetilde{\mathrm{Y}}_i}$ represents the Dirac measure at $(\mathrm{Z}_i, \widetilde{\mathrm{Y}}_i)$. With this setup, we proceed to design a robust loss function over $\widehat{P}^\dagger$ that is resilient to the heterogeneous label noise in $\widetilde{\mathrm{Y}}$, while admitting a residual-based reformulation for explicit noise modeling.

4.1.1 Robust Loss Family

Label noise in regression is heterogeneous in magnitude: small-to-moderate deviations are addressable at the instance level, while large deviations, which often arise from outlying or heavily corrupted annotations, render exact label values unreliable, such that their influence on training should be limited rather than fully absorbed by the model. To this end, we introduce a piecewise loss family that applies quadratic sensitivity within a bounded range while controlling the growth of penalties for large deviations:

$$\ell_\tau(\nu) = \begin{cases} \frac{1}{2}\nu^2, & \text{if } |\nu| \le \tau, \\ \frac{\tau}{\ell'(\tau)}\{\ell(|\nu|) - \ell(\tau)\} + \frac{\tau^2}{2}, & \text{if } |\nu| > \tau, \end{cases} \tag{3}$$

where $\tau > 0$ is a capacity threshold, $\ell(\cdot) : [\tau, \infty) \to \mathbb{R}_{\ge 0}$ governs large-residual penalization, and the factor $\frac{\tau}{\ell'(\tau)}$ ensures continuous differentiability at $|\nu| = \tau$, with matching value $\frac{\tau^2}{2}$ and derivative τ from both sides. The loss ℓ_τ reduces to MSE when $|\nu| \le \tau$, preserving sensitivity to local prediction errors. For $|\nu| > \tau$, its growth is governed by $\ell(\cdot)$, ensuring sub-quadratic penalization

of large residuals and preventing large deviations from being overfitted, leaving their treatment to complementary structural components. Let $\widehat{\theta}^{\tau}_{\text{reg}}$ denote the minimizer of the following empirical risk:

$$\Re^{\tau}_{\text{reg}}(\theta_{\text{reg}}) \triangleq \frac{1}{N}\sum_{i=1}^{N} \ell_{\tau}(\langle \mathrm{Z}_i, \theta_{\text{reg}}\rangle - \widetilde{\mathrm{Y}}_i). \tag{4}$$

We next identify conditions on $\ell(\cdot)$ under which $\widehat{\theta}^{\tau}_{\text{reg}}$ admits a generalization guarantee.

Assumption 4.1. Assume that $\ell(\cdot)$ is convex, twice differentiable, and satisfies $\ell'(t)>0$ on $[\tau,\infty)$. Moreover, $\ell'(t)/t$ is non-increasing on $[\tau,\infty)$, or equivalently, $t\ell''(t)-\ell'(t)\leq 0$ for all $t\geq\tau$, and $\ell'(t)\leq t^{\kappa}$ for all $t\geq\tau$, where $\kappa\geq 0$ is the constant specified in Condition A.2 (iii)(b).

Intuitively, the non-increasing $\ell'(t)/t$ condition ensures that the effective penalization per unit residual decreases as residual magnitude grows, providing robustness to large deviations without discarding them entirely. Common choices satisfying this assumption include $\ell(t)=t^q$ for $1\leq q<2$ and $t\geq\tau$. Under mild regularity conditions, we establish the following generalization guarantee for $\widehat{\theta}^{\tau}_{\text{reg}}$. Denote $\Sigma\triangleq\mathbb{E}(\mathrm{Z}\mathrm{Z}^{\top})$, and for any $u\in\mathbb{R}^d$, let $\|u\|_{\Sigma}\triangleq\|\Sigma^{1/2}u\|_2$.

Theorem 4.2. *Assume that the regularity conditions in Appendix A.2 hold, and that the function $\ell(\cdot)$ in* (3) *satisfies Assumption 4.1. Suppose that $d>10$, and $\tau\geq\mathscr{C}_1(\sqrt{d}+\sigma)$, where $\sigma>0$ is defined in Condition (ii) of Appendix A.2 and $\mathscr{C}_1>0$ is a constant. Furthermore, for any $s>0$, assume that $\max\Big(\tau,\frac{\tau}{\ell'(\tau)}\Big)\leq\sigma\sqrt{\frac{N}{d+s}}$, and $N>\mathscr{C}_2(d+s)$ for some constant $\mathscr{C}_2>0$. Then,*

(i) with probability at least $1-5e^{-s}$, $\|\widehat{\theta}^{\tau}_{reg}-\theta^{*}_{reg}\|_{\Sigma}\leq\mathscr{C}^{\dagger}\sqrt{\frac{d+s}{N}}$ for some constant $\mathscr{C}^{\dagger}>0$;

(ii) with probability at least $1-7e^{-s}$, for any future input X_{new} with feature $\mathrm{Z}_{new}\triangleq\varphi(\mathrm{X}_{new})$, the associated prediction error satisfies: $err\triangleq\Big|\langle\mathrm{Z}_{new},\widehat{\theta}^{\tau}_{reg}\rangle-\langle\mathrm{Z}_{new},\theta^{*}_{reg}\rangle\Big|\leq\mathscr{C}^{\ddagger}\sqrt{\frac{ds+s^2}{N}}$ for some constant $\mathscr{C}^{\ddagger}>0$.

Remark 4.3. A distributionally robust extension of Theorem 4.2, which accounts for distributional uncertainty induced by noisy labels, is presented in Theorem A.15 in Appendix A.3. Theoretical limitations and future directions are discussed in Appendix C.

Theorem 4.2 establishes non-asymptotic guarantees for both parameter estimation and prediction under heterogeneous label noise. In particular, the parameter estimation error is bounded at the rate $\mathcal{O}(\sqrt{(d+s)/N})$, while the prediction error admits a bound of order $\mathcal{O}(\sqrt{(ds+s^2)/N})$. Beyond these statistical guarantees, for any $\ell(\cdot)$ satisfying Assumption 4.1, the corresponding piecewise loss ℓ_{τ} in (3) admits an infimal convolution representation, which provides a principled bridge between the loss design and the fine-grained residual modeling mechanism in our framework, as detailed in Section 4.1.2.

4.1.2 Residual-Based Reformulation

Building on the piecewise loss $\ell_{\tau}(\cdot)$, we derive an equivalent reformulation of the empirical objective $\Re^{\tau}_{\text{reg}}$ in (4) that explicitly separates instance-specific label deviations from the predictive signal, thus providing the theoretical foundation for the residual head design developed in Section 4.1.3.

Proposition 4.4. *Let $\ell:[\tau,\infty)\to\mathbb{R}_{\geq 0}$ be convex and differentiable, with $\ell'(\tau)>0$. Assume that $\ell'(t)/t$ is strictly decreasing on $[\tau,\infty)$. Then $\ell_{\tau}(\cdot)$ in* (3) *is convex and admits the infimal convolution representation*

$$\ell_{\tau}(\nu)=\inf_{\xi\in\mathbb{R}}\left\{\frac{1}{2}\xi^2+\phi_{\tau}(\nu-\xi)\right\}, \tag{5}$$

where $\phi_{\tau}:\mathbb{R}\to\mathbb{R}_{\geq 0}$ is defined via the following parametric characterization: for any $t\geq\tau$, let $u=t-\frac{\tau\ell'(t)}{\ell'(\tau)}$; then

$$\phi_{\tau}(u)=\frac{\tau}{\ell'(\tau)}\big\{\ell(t)-\ell(\tau)\big\}+\frac{\tau^2}{2}-\frac{1}{2}\left\{\frac{\tau\ell'(t)}{\ell'(\tau)}\right\}^2. \tag{6}$$

The mapping $t \mapsto u$ is continuous and strictly increasing from $[\tau, \infty)$ onto $[0, \infty)$, so that each $u \geq 0$ corresponds to a unique $t \geq \tau$. For $u < 0$, define $\phi_\tau(u) = \phi_\tau(-u)$. Moreover, ϕ_τ is convex, even, and satisfies $\phi_\tau(0) = 0$.

Remark 4.5. When $\ell(t) = t$ for $t \geq \tau$, the piecewise loss ℓ_τ in (3) reduces to the Huber loss with threshold τ. In this case, $\phi_\tau(u) = \tau|u|$, yielding $\ell_\tau(\nu) = \inf_{\xi \in \mathbb{R}} \left\{ \frac{1}{2}\xi^2 + \tau|\nu - \xi| \right\}$.

With Proposition 4.4, minimizing the empirical risk $\Re^\tau_{\text{reg}}(\theta_{\text{reg}})$ in (4) over θ_{reg} becomes equivalent to the following joint minimization over θ_{reg} and $\boldsymbol{\xi}$:

$$\Re^\tau_{\text{reg}}(\theta_{\text{reg}}, \boldsymbol{\xi}) \triangleq \frac{1}{N} \Bigg\{ \sum_{i=1}^{N} \underbrace{\frac{1}{2}\xi_i^2}_{\text{residual capacity penalty}} + \underbrace{\phi_\tau\left(\widetilde{\mathrm{Y}}_i - \langle \mathrm{Z}_i, \theta_{\text{reg}} \rangle - \xi_i\right)}_{\text{explicit residual learning}} \Bigg\}, \tag{7}$$

where $\boldsymbol{\xi} \triangleq \{\xi_i\}_{i=1}^N$, and each auxiliary variable $\xi_i \in \mathbb{R}$ captures the instance-specific label deviations, with the penalty $\frac{1}{2}\xi_i^2$ discouraging large residuals and preventing absorption of predictive signals. This decomposition allows the regression model to focus on the underlying predictive mapping while delegating such deviations to ξ_i, preventing noise from contaminating global model parameters.

4.1.3 Residual Modeling Design

The reformulation in (7) naturally suggests modeling $\{\xi_i\}$ as instance-specific residuals, but directly optimizing over free variables risks degenerate solutions in which residuals absorb prediction signals, leaving the regression model under-trained. To prevent this, we parameterize the residual variables through a learnable function on the feature representation. For the algorithmic implementation, we extend the fixed feature representation considered in Sections 4.1.1–4.1.2 to a learnable feature mapping $\varphi(\cdot; \theta_{\text{f}}) : \mathcal{X} \to \mathcal{Z}$ parameterized by θ_{f}. Specifically, we introduce a lightweight residual head $\hbar_{\text{res}}(\cdot; \theta_{\text{res}})$ built on top of the feature extractor $\varphi(\cdot; \theta_{\text{f}})$, with θ_{res} denoting the parameters of the residual head; and define

$$\xi_i(\theta_{\text{res}}) \triangleq \tau \cdot \tanh\left(\hbar_{\text{res}}(\mathrm{Z}_i; \theta_{\text{res}})\right), \quad \mathrm{Z}_i \triangleq \varphi(\mathrm{X}_i; \theta_{\text{f}}), \tag{8}$$

which enforces the capacity constraint $\xi_i \in (-\tau, \tau)$. Under this parameterization, residual correction is tied to the learned feature representation rather than treated as independent degrees of freedom, ensuring that residual correction is consistent with the learned structure of the data.

Motivated by the residual decomposition in (7), we train the residual head using the following pseudo-label-guided objective:

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{2}\xi_i^2(\theta_{\text{res}}) + \phi_\tau\left((\widetilde{\mathrm{Y}}_i - \mathrm{Y}_i^*) - \xi_i(\theta_{\text{res}})\right) \right\}, \tag{9}$$

where Y_i^* denotes a robust pseudo label whose construction is specified in (B.5) of Appendix B.2. The residual head targets the deviation $\widetilde{\mathrm{Y}}_i - \mathrm{Y}_i^*$, decoupling residual correction from the predictive signal. The residual parameters θ_{res} are updated in an alternating manner with other model parameters, which prevents the residual head from co-adapting with the regression model and ensures that ξ_i captures only instance-specific deviations rather than replacing the role of the regression model.

Finally, rather than fixing τ as a hyperparameter, we adapt it each epoch as a scaled median absolute deviation (MAD) [17] of the current prediction residuals $\{\widetilde{\mathrm{Y}}_i - \mathrm{Y}_i^*\}_{i=1}^N$ (detailed in Appendix B.2). This adaptive τ defines a principled boundary between the two granularities: deviations within $(-\tau, \tau)$ are handled through instance-level residual refinement, while larger deviations, which cannot be reliably captured at the instance level, are instead addressed by the coarse-grained structural mechanism introduced in Section 4.2, as illustrated in Figure 1.

4.2 Coarse-Grained Structuring: Contrastive Regularization across Bins

For large deviations, where exact label values become unreliable, direct supervision may mislead the model and induce structural distortion in the feature space: representations no longer reflect label proximity, and the global geometry degrades. While the fine-grained module performs instance-level correction within $(-\tau, \tau)$ under label supervision, it does not directly enforce a globally consistent, order-preserving structure aligned with label proximity. Accordingly, we introduce a coarse-grained structuring mechanism that operates at the distribution level, enforcing an order-preserving geometry across bins of the label space.

4.2.1 Discretization of Label Space

We partition the label space $\mathcal{Y}$ into K equal-width bins $\mathcal{Y} = \cup_{k=1}^{K} \mathcal{Y}_k$ (detailed in Appendix B.2). For $k \in [K]$ with $[K] \triangleq \{1, \ldots, K\}$, let $\mathcal{I}_k$ index samples whose labels fall into bin k, where after a warm-up phase bin assignment is based on the robust pseudo labels $\{\mathrm{Y}_i^*\}_{i=1}^N$ (detailed in Appendix B.2) rather than the original noisy labels. Define $\mathcal{D}_k \triangleq \{(\mathrm{X}_i, \widetilde{\mathrm{Y}}_i) : i \in \mathcal{I}_k\}$. Based on this discretization, we introduce a histogram head $\hbar_{\text{hist}}(\cdot; \theta_{\text{hist}})$ parameterized by θ_{hist} on top of the feature extractor $\varphi(\cdot; \theta_{\text{f}})$, producing logits $s_i = \hbar_{\text{hist}}(\varphi(\mathrm{X}_i; \theta_{\text{f}}); \theta_{\text{hist}}) \in \mathbb{R}^K$ and bin softmax probabilities $p_i = \mathrm{softmax}(s_i) \in \Delta^{K-1}$ over the K-simplex Δ^{K-1}.

The bin structure serves two purposes. First, following prior work [13, 15], classification-based objectives encourage greater feature diversity than regression-only training; accordingly, we include a cross-entropy loss over bin assignments to mitigate representation collapse under noisy supervision:

$$\mathfrak{R}_{\text{ce}}(\theta_{\text{hist}}, \theta_{\text{f}}) = -\frac{1}{N} \sum_{k \in [K]} \sum_{i \in \mathcal{I}_k} \log p_i[k], \tag{10}$$

Second, and more centrally, the bin structure provides the foundation for the cross-bin contrastive regularization below, which constitutes the core coarse-grained structuring mechanism.

4.2.2 Order-Preserving Cross-Bin Contrastive Objective

We construct contrastive relations at the level of bin-wise feature distributions. For each bin k, define the empirical feature distribution $\widehat{P}_z^{(k)}(\theta_{\text{f}}) = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \delta_{\mathrm{Z}_i(\theta_{\text{f}})}$, where $\mathrm{Z}_i(\theta_{\text{f}}) \triangleq \varphi(\mathrm{X}_i; \theta_{\text{f}})$, $|\mathcal{I}_k|$ denotes the cardinality of $\mathcal{I}_k$, and δ_v denotes the Dirac measure concentrated at v. For each $k \in [K]$, we form a triplet $\big(\widehat{P}_z^{(k)}(\theta_{\text{f}}), \widehat{P}_z^{(k_{\text{pos}})}(\theta_{\text{f}}), \widehat{P}_z^{(k_{\text{neg}})}(\theta_{\text{f}})\big)$, where k_{pos} and k_{neg} denote the indices of nearby and distant bins, respectively, satisfying $|k - k_{\text{pos}}| < |k - k_{\text{neg}}|$.

Next, we define positive and negative bin neighborhoods based on label proximity. Specifically, for a positive integer R and $k \in [K]$, let $\mathcal{N}_k^+ \triangleq \{k-R, \ldots, k-1, k+1, \ldots, k+R\} \cap [K]$ index the nearest neighboring bins of bin k, and $\mathcal{N}_k^- \triangleq \{k - 2R, \ldots, k - (R+1), k + (R+1), \ldots, k + 2R\} \cap [K]$ denote the indices of farther bins. The contrastive losses are then computed as weighted averages:

$$\mathfrak{R}_{\text{pos}} \triangleq \frac{1}{K} \sum_{k=1}^{K} \sum_{k' \in \mathcal{N}_k^+} \frac{1/|k - k'|}{\sum_{j \in \mathcal{N}_k^+} 1/|k - j|} \, \mathscr{d}\big(\widehat{P}_z^{(k)}, \widehat{P}_z^{(k')}\big); \tag{11}$$

$$\mathfrak{R}_{\text{neg}} \triangleq -\frac{1}{K} \sum_{k=1}^{K} \sum_{k' \in \mathcal{N}_k^-} \frac{(k - k')^2}{\sum_{j \in \mathcal{N}_k^-} (k - j)^2} \, \mathscr{d}\big(\widehat{P}_z^{(k)}, \widehat{P}_z^{(k')}\big). \tag{12}$$

where θ_{f}-dependence of $\mathfrak{R}_{\text{pos}}$, $\mathfrak{R}_{\text{neg}}$, $\widehat{P}_z^{(k)}$, and $\widehat{P}_z^{(k')}$ is omitted temporarily for brevity, and $\mathscr{d}(P, Q)$ denotes a discrepancy measure between distributions P and Q. Here, the positive weights emphasize closer neighbors, while the negative weights penalize more distant bin pairs proportionally to their label separation. We set $R = 1$ by default. In our implementation, we employ an unbalanced optimal transport (UOT) cost (Definition A.2) under a Gaussian approximation [18] as the discrepancy measure, providing further robustness to bin misassignments induced by label noise.

This objective enforces an order-preserving feature geometry: representations from nearby bins are pulled together to preserve label continuity, while distant bins are pushed apart to maintain discriminability. Unlike instance-level residual refinement, this coarse-grained structuring mechanism directly targets the structural corruption induced by large-magnitude noise, namely the deviations that lie beyond the capacity threshold τ and fall outside the scope of fine-grained correction.

4.3 Unified Framework: Bridging across Granularities

Let $\vartheta \triangleq (\theta_{\text{reg}}^\top, \theta_{\text{hist}}^\top, \theta_{\text{res}}^\top, \theta_{\text{f}}^\top)^\top$ denote all trainable parameters. Combining (7), (10), (11) and (12), and suppressing parameter dependence for notational simplicity, the joint training objective is

$$\mathcal{L}(\vartheta) = \lambda_{\text{reg}} \mathfrak{R}_{\text{reg}}^\tau + \lambda_{\text{ce}} \mathfrak{R}_{\text{ce}} + \lambda_{\text{pos}} \mathfrak{R}_{\text{pos}} + \lambda_{\text{neg}} \mathfrak{R}_{\text{neg}}, \tag{13}$$

where λ_{reg}, λ_{ce}, λ_{pos}, and λ_{neg} are positive hyperparameters balancing the respective components. In practice, we fix $\lambda_{\text{ce}} = 1.0$ and the ratio $\lambda_{\text{neg}}/\lambda_{\text{pos}} = 0.5$, reducing the tuning to two effective hyperparameters: λ_{reg} and λ_{pos}. Training details and Algorithm 1 are deferred to Appendix B.2.

The key component bridging the two granularities is the histogram head, which generates robust pseudo labels $\{\mathrm{Y}_i^*\}$ from the shared feature representations and couples the two modules through a unified supervisory signal. Each pseudo label is initialized from the observed noisy label $\widetilde{\mathrm{Y}}_i$ and updated each epoch via exponential moving average (EMA) of the histogram head's weighted bin-center predictions, retaining the smallest set of highest-probability bins whose cumulative probability mass reaches at least 68% (motivated by one-standard-deviation Gaussian coverage), with confidence computed as $1 - K^*/K$, where K^* is the number of retained bins; full details are given in (B.2)-(B.6) of Appendix B.2.

The pseudo labels $\{\mathrm{Y}_i^*\}$ from the histogram head couple the two modules through a shared supervisory signal. In the fine-grained module, they provide residual targets $\widetilde{\mathrm{Y}}_i - \mathrm{Y}_i^*$ in (9), decoupling residual estimation from the noisy label directly. In the coarse-grained module, they stabilize bin assignments under noisy supervision, improving the reliability of the contrastive objective. Through these roles, the two modules jointly produce structured and refined feature representations: the coarse-grained module enforces global order in the feature space, establishing a well-organized geometric structure aligned with label proximity, while fine-grained refinement performs precise instance-level correction within this structured space. The resulting representations, in turn, improve the quality of histogram head predictions, yielding more reliable pseudo labels in subsequent epochs. This closed-loop interaction between pseudo-label estimation and feature learning allows each component to progressively reinforce the others, enabling coordinated handling of both global structural distortion and instance-level label noise throughout training.

5 Experiment Details

5.1 Experimental Setup

Evaluation Metrics. We evaluate model performance using three standard regression metrics: mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2), where lower values of MAE and RMSE indicate better predictive accuracy, and higher values of R^2 indicate a higher proportion of variance explained by the model. In the reported tables, we show the desired direction using "↓" for metrics to minimize and "↑" for metrics to maximize. The detailed definitions and computation of each metric are deferred to Appendix B.1.

Datasets. We conduct experiments on four datasets spanning three different regression tasks: **Age estimation** on UTKFace [19] and APPA-REAL [20], **Head-pose estimation** on Biwi Kinect [21], and **Bone-age assessment** on RSNA [22] combined with DHA [23]. For UTKFace and Biwi Kinect, we simulate label noise by adding Gaussian perturbations with varying maximal standard deviations std_{max}. Here, std_{max} controls the magnitude of the injected noise during generation, rather than representing the actual standard deviation of the resulting noisy labels; higher values correspond to stronger noise. In our experiments, we consider three levels, $\text{std}_{\text{max}} \in \{30, 60, 90\}$. In contrast, APPA-REAL and the Bone-Age datasets contain inherent, real-world label noise. Additional details on dataset preprocessing and noise simulation are provided in Appendix B.1.

Baselines. We consider the following baselines for regression tasks: (i) Vanilla, which directly trains with MSE; (ii) SuperLoss [7]; (iii) RDI [5]; (iv) AUX [5]; and (v) C-Mixup [6]. We also include several classical methods originally designed for noisy-label classification that can be adapted to regression: (1) CDR [3], (2) CoTeaching [2], (3) CNLCU-S [8], (4) CNLCU-H [8], (5) Selfie [4], and (6) CoSelfie [4]. Detailed descriptions of all methods are provided in Appendix B.1. For fair comparison, all baselines are implemented in two ways: (a) using the original regression loss directly, and (b) with a histogram head following the same binning strategy as GRAIN. In the reported tables, results from the latter implementation are indicated by "BIN". Baselines that simultaneously train two models are further indicated by the superscript symbol "†". For both GRAIN and all baselines, we adopt standard (basic) data augmentation during training, including random cropping, horizontal flipping, and rotation. Results obtained with slightly stronger data augmentations, which further add color jitter and mixed image transformations, are also reported and denoted as "GRAIN-AUG".

Table 1: Mean $\pm$ standard deviation on UTKFace under simulated Gaussian label noise with corruption levels controlled by $\mathrm{std}_{\max}$, and on Bone-Age. The best baseline for each metric is highlighted in blue, and results from GRAIN and GRAIN-AUG are presented in **boldface**.

Method	UTKFace ($\mathrm{std}_{\max}$= 30)			UTKFace ($\mathrm{std}_{\max}$= 60)			UTKFace ($\mathrm{std}_{\max}$= 90)			Bone-Age		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
Vanilla	$8.13_{\pm 0.17}$	$11.20_{\pm 0.12}$	$0.70_{\pm 0.01}$	$9.94_{\pm 0.19}$	$13.07_{\pm 0.22}$	$0.59_{\pm 0.01}$	$10.25_{\pm 0.29}$	$13.43_{\pm 0.31}$	$0.56_{\pm 0.02}$	$13.56_{\pm 0.25}$	$17.44_{\pm 0.36}$	$0.91_{\pm 0.00}$
SuperLoss	$7.69_{\pm 0.19}$	$11.43_{\pm 0.41}$	$0.68_{\pm 0.02}$	$8.18_{\pm 0.12}$	$11.95_{\pm 0.32}$	$0.66_{\pm 0.02}$	$8.23_{\pm 0.24}$	$11.97_{\pm 0.33}$	$0.65_{\pm 0.02}$	$13.59_{\pm 0.11}$	$17.33_{\pm 0.18}$	$0.91_{\pm 0.00}$
RDI	$8.09_{\pm 0.11}$	$11.31_{\pm 0.21}$	$0.69_{\pm 0.01}$	$9.71_{\pm 0.29}$	$12.86_{\pm 0.23}$	$0.60_{\pm 0.01}$	$10.03_{\pm 0.26}$	$13.31_{\pm 0.32}$	$0.57_{\pm 0.02}$	$13.26_{\pm 0.34}$	$17.23_{\pm 0.49}$	$0.91_{\pm 0.01}$
AUX	$8.19_{\pm 0.16}$	$11.40_{\pm 0.31}$	$0.69_{\pm 0.02}$	$9.74_{\pm 0.24}$	$12.90_{\pm 0.16}$	$0.60_{\pm 0.01}$	$10.23_{\pm 0.36}$	$13.51_{\pm 0.29}$	$0.56_{\pm 0.02}$	$13.35_{\pm 0.11}$	$17.21_{\pm 0.13}$	$0.91_{\pm 0.00}$
C-Mixup	$7.56_{\pm 0.16}$	$10.40_{\pm 0.24}$	$0.74_{\pm 0.01}$	$9.69_{\pm 0.15}$	$13.07_{\pm 0.32}$	$0.59_{\pm 0.02}$	$10.11_{\pm 0.24}$	$13.45_{\pm 0.30}$	$0.56_{\pm 0.02}$	$13.51_{\pm 0.17}$	$17.52_{\pm 0.23}$	$0.91_{\pm 0.00}$
CDR	$8.21_{\pm 0.15}$	$11.49_{\pm 0.07}$	$0.68_{\pm 0.00}$	$9.88_{\pm 0.15}$	$13.06_{\pm 0.23}$	$0.59_{\pm 0.02}$	$10.39_{\pm 0.25}$	$13.71_{\pm 0.18}$	$0.55_{\pm 0.01}$	$13.34_{\pm 0.14}$	$17.34_{\pm 0.16}$	$0.91_{\pm 0.00}$
Selfie	$8.09_{\pm 0.06}$	$12.32_{\pm 0.24}$	$0.63_{\pm 0.01}$	$8.30_{\pm 0.12}$	$12.70_{\pm 0.29}$	$0.61_{\pm 0.02}$	$8.38_{\pm 0.29}$	$12.77_{\pm 0.45}$	$0.61_{\pm 0.03}$	$15.43_{\pm 0.76}$	$19.65_{\pm 0.86}$	$0.88_{\pm 0.01}$
CoTeaching †	$7.84_{\pm 0.16}$	$11.95_{\pm 0.37}$	$0.66_{\pm 0.02}$	$7.98_{\pm 0.09}$	$12.18_{\pm 0.28}$	$0.64_{\pm 0.02}$	$8.28_{\pm 0.29}$	$12.71_{\pm 0.49}$	$0.61_{\pm 0.03}$	$14.39_{\pm 0.41}$	$18.36_{\pm 0.70}$	$0.90_{\pm 0.01}$
CNLCU-S †	$8.16_{\pm 0.15}$	$12.64_{\pm 0.35}$	$0.61_{\pm 0.02}$	$8.17_{\pm 0.22}$	$12.66_{\pm 0.45}$	$0.61_{\pm 0.03}$	$8.48_{\pm 0.32}$	$13.04_{\pm 0.59}$	$0.59_{\pm 0.04}$	$15.97_{\pm 0.53}$	$20.54_{\pm 0.71}$	$0.87_{\pm 0.01}$
CNLCU-H †	$8.30_{\pm 0.18}$	$12.83_{\pm 0.37}$	$0.60_{\pm 0.02}$	$8.31_{\pm 0.15}$	$12.86_{\pm 0.33}$	$0.60_{\pm 0.02}$	$8.49_{\pm 0.35}$	$12.99_{\pm 0.58}$	$0.59_{\pm 0.04}$	$15.83_{\pm 0.68}$	$20.12_{\pm 0.94}$	$0.88_{\pm 0.01}$
CoSelfie †	$7.84_{\pm 0.15}$	$11.87_{\pm 0.38}$	$0.66_{\pm 0.02}$	$8.03_{\pm 0.22}$	$12.30_{\pm 0.52}$	$0.63_{\pm 0.03}$	$8.21_{\pm 0.22}$	$12.47_{\pm 0.34}$	$0.62_{\pm 0.02}$	$15.50_{\pm 0.46}$	$19.70_{\pm 0.46}$	$0.88_{\pm 0.01}$
Vanilla (BIN)	$7.79_{\pm 0.09}$	$11.87_{\pm 0.12}$	$0.66_{\pm 0.01}$	$8.44_{\pm 0.17}$	$13.12_{\pm 0.30}$	$0.58_{\pm 0.02}$	$8.46_{\pm 0.28}$	$13.06_{\pm 0.44}$	$0.59_{\pm 0.03}$	$13.31_{\pm 0.22}$	$17.43_{\pm 0.53}$	$0.91_{\pm 0.01}$
SuperLoss (BIN)	$7.61_{\pm 0.19}$	$11.35_{\pm 0.27}$	$0.69_{\pm 0.01}$	$7.90_{\pm 0.19}$	$11.61_{\pm 0.31}$	$0.67_{\pm 0.02}$	$7.95_{\pm 0.19}$	$11.65_{\pm 0.26}$	$0.67_{\pm 0.01}$	$13.55_{\pm 0.33}$	$17.41_{\pm 0.33}$	$0.91_{\pm 0.00}$
RDI (BIN)	$7.84_{\pm 0.12}$	$11.89_{\pm 0.17}$	$0.66_{\pm 0.01}$	$8.41_{\pm 0.10}$	$12.88_{\pm 0.27}$	$0.60_{\pm 0.02}$	$8.46_{\pm 0.21}$	$13.29_{\pm 0.45}$	$0.57_{\pm 0.03}$	$13.18_{\pm 0.15}$	$16.84_{\pm 0.33}$	$0.91_{\pm 0.00}$
AUX (BIN)	$7.70_{\pm 0.09}$	$11.68_{\pm 0.12}$	$0.67_{\pm 0.01}$	$8.35_{\pm 0.10}$	$12.73_{\pm 0.08}$	$0.61_{\pm 0.01}$	$8.45_{\pm 0.18}$	$13.20_{\pm 0.25}$	$0.58_{\pm 0.02}$	$13.27_{\pm 0.31}$	$17.16_{\pm 0.68}$	$0.91_{\pm 0.01}$
C-Mixup (BIN)	$7.44_{\pm 0.25}$	$10.40_{\pm 0.36}$	$0.74_{\pm 0.02}$	$8.29_{\pm 0.14}$	$12.84_{\pm 0.32}$	$0.60_{\pm 0.02}$	$8.49_{\pm 0.21}$	$13.24_{\pm 0.29}$	$0.58_{\pm 0.02}$	$13.12_{\pm 0.54}$	$17.45_{\pm 1.04}$	$0.91_{\pm 0.01}$
CDR (BIN)	$7.74_{\pm 0.15}$	$10.82_{\pm 0.25}$	$0.72_{\pm 0.01}$	$8.56_{\pm 0.28}$	$13.48_{\pm 0.67}$	$0.56_{\pm 0.04}$	$8.58_{\pm 0.27}$	$13.16_{\pm 0.33}$	$0.58_{\pm 0.02}$	$13.45_{\pm 0.29}$	$17.52_{\pm 0.84}$	$0.91_{\pm 0.01}$
Selfie (BIN)	$8.20_{\pm 0.36}$	$12.50_{\pm 0.66}$	$0.62_{\pm 0.04}$	$8.44_{\pm 0.47}$	$12.79_{\pm 0.72}$	$0.60_{\pm 0.05}$	$9.07_{\pm 0.44}$	$13.83_{\pm 0.48}$	$0.54_{\pm 0.03}$	$16.35_{\pm 1.13}$	$20.81_{\pm 1.56}$	$0.87_{\pm 0.02}$
CoTeaching (BIN) †	$8.22_{\pm 0.33}$	$12.66_{\pm 0.61}$	$0.61_{\pm 0.04}$	$8.29_{\pm 0.32}$	$12.66_{\pm 0.64}$	$0.61_{\pm 0.04}$	$8.75_{\pm 0.47}$	$13.26_{\pm 0.53}$	$0.58_{\pm 0.03}$	$14.58_{\pm 0.47}$	$18.68_{\pm 0.75}$	$0.89_{\pm 0.01}$
CNLCU-S (BIN) †	$7.89_{\pm 0.21}$	$12.19_{\pm 0.44}$	$0.64_{\pm 0.03}$	$8.29_{\pm 0.60}$	$12.84_{\pm 1.27}$	$0.60_{\pm 0.08}$	$8.34_{\pm 0.39}$	$13.04_{\pm 0.59}$	$0.59_{\pm 0.04}$	$15.96_{\pm 0.79}$	$20.22_{\pm 1.01}$	$0.88_{\pm 0.01}$
CNLCU-H (BIN) †	$7.91_{\pm 0.18}$	$12.24_{\pm 0.37}$	$0.64_{\pm 0.02}$	$8.45_{\pm 0.52}$	$13.13_{\pm 0.91}$	$0.58_{\pm 0.06}$	$8.58_{\pm 0.25}$	$13.29_{\pm 0.38}$	$0.57_{\pm 0.02}$	$17.39_{\pm 0.97}$	$22.29_{\pm 1.37}$	$0.85_{\pm 0.02}$
CoSelfie (BIN) †	$8.29_{\pm 0.27}$	$12.61_{\pm 0.31}$	$0.62_{\pm 0.02}$	$8.46_{\pm 0.86}$	$12.90_{\pm 1.33}$	$0.59_{\pm 0.09}$	$8.60_{\pm 0.73}$	$13.06_{\pm 1.01}$	$0.59_{\pm 0.06}$	$14.44_{\pm 0.49}$	$18.62_{\pm 0.56}$	$0.89_{\pm 0.01}$
GRAIN	$\mathbf{7.01}_{\pm 0.10}$	$\mathbf{10.28}_{\pm 0.24}$	$\mathbf{0.74}_{\pm 0.01}$	$\mathbf{7.54}_{\pm 0.14}$	$\mathbf{11.10}_{\pm 0.33}$	$\mathbf{0.70}_{\pm 0.02}$	$\mathbf{7.65}_{\pm 0.12}$	$\mathbf{11.28}_{\pm 0.22}$	$\mathbf{0.70}_{\pm 0.01}$	$\mathbf{12.85}_{\pm 0.26}$	$\mathbf{16.55}_{\pm 0.32}$	$\mathbf{0.92}_{\pm 0.00}$
GRAIN-AUG	$\mathbf{6.81}_{\pm 0.13}$	$\mathbf{9.98}_{\pm 0.28}$	$\mathbf{0.76}_{\pm 0.01}$	$\mathbf{7.30}_{\pm 0.14}$	$\mathbf{10.74}_{\pm 0.28}$	$\mathbf{0.72}_{\pm 0.01}$	$\mathbf{7.35}_{\pm 0.22}$	$\mathbf{10.75}_{\pm 0.36}$	$\mathbf{0.72}_{\pm 0.02}$	$\mathbf{12.80}_{\pm 0.22}$	$\mathbf{16.40}_{\pm 0.29}$	$\mathbf{0.92}_{\pm 0.00}$

Training Details. For all datasets except Bone-Age, we use a ResNet-50 backbone [24]; for the Bone-Age dataset, we employ DenseNet-121 [25]. The initial learning rate is taken as 5×10^{-4} or 10^{-3}, and a cosine learning rate scheduler is applied. For all methods, each noisy dataset is split into training and validation sets with a $4:1$ ratio, and model selection is performed based on validation MAE. All experiments are repeated with five random seeds (1–5).

5.2 Main Results

Comparison with Baseline Methods. Table 1 summarizes the results on UTKFace under simulated Gaussian label noise with different corruption levels, and on Bone-Age dataset with real-world label noise, evaluated using three metrics. Across nearly all noise levels and metrics, GRAIN and GRAIN-AUG consistently demonstrate the best performance, showing strong robustness to both synthetic and real-world label noise. Additional results on the Biwi Kinect dataset with synthetic Gaussian noise are reported in Tables B.3-B.5, and results on APPA-REAL with real-world label noise are reported in Table B.2, all appearing in Appendix B.3; they all exhibit the same patterns.

Ablation Studies. We conduct an ablation study to isolate the contributions of key components in GRAIN. Specifically, we consider: (i) adaptive binning procedure in Appendix B.2 with the CE loss ("CE"); (ii) direct training with the Huber loss ("Huber Loss"); (iii) training the residual head h_{res} using the regression head output in place of Y_i^* in (B.5) ("GRAIN using $\mathrm{Y}_{\text{reg}}^*$"); (iv) GRAIN without fine-grained residual learning ("GRAIN w.o. Fine-Grained"); (v) computing Y_i^* in (B.5) using the full classification softmax output instead of the top 68% predictions ("GRAIN using $\mathrm{Y}_{\text{full}}^*$"); (vi) GRAIN without coarse-grained cross-bin contrastive regularization ("GRAIN w.o. Coarse-Grained"); (vii) GRAIN without the negative term $\Re_{\text{neg}}$ ("GRAIN w.o. $\Re_{\text{neg}}$"); and (viii) the full GRAIN model ("GRAIN"). Figure 3 reports the results on UTKFace under simulated Gaussian label noise at varying corruption levels, highlighting the individual contributions of each related component.

Hyperparameter Sensitivity Analysis. We further analyze the sensitivity of GRAIN to different loss coefficients. Since the magnitude of the CE loss varies less than that of the regression loss during training, we fix the CE loss weight to $\lambda_{\text{ce}} = 1.0$ in all experiments, and treat the regression loss weight λ_{reg} as a hyperparameter. Figure 4 reports the performance under different values of λ_{reg} on UTKFace with simulated Gaussian label noise at varying corruption levels. The red dashed line indicates the performance at $\lambda_{\text{reg}} = 0$, corresponding to the removal of the regression loss; and the black dashed line denotes the best-performing baseline method. The results show that GRAIN achieves consistently strong performance within a reasonable range of λ_{reg}. Additional sensitivity analyses on λ_{pos}, the ratio $\lambda_{\text{neg}}/\lambda_{\text{pos}}$, the number of contrastive pairs, and the number of bins are provided in Figures B.1, B.2, B.3, and Figures B.4-B.5 in Appendix B.3. These results show that GRAIN remains robust across a wide range of hyperparameter choices under varying noise levels.

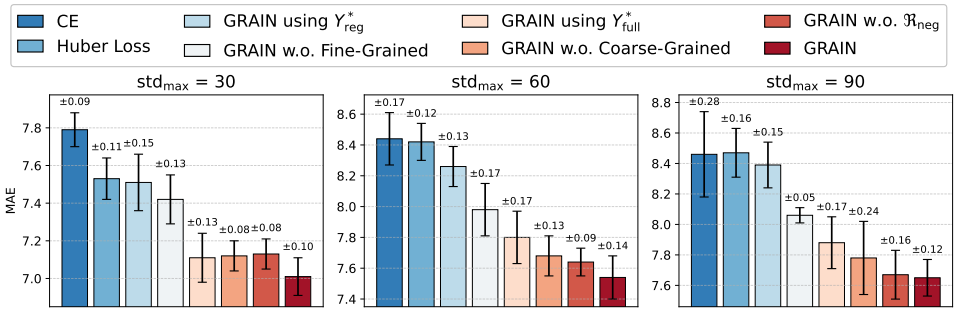


Figure 3: Ablation study results.

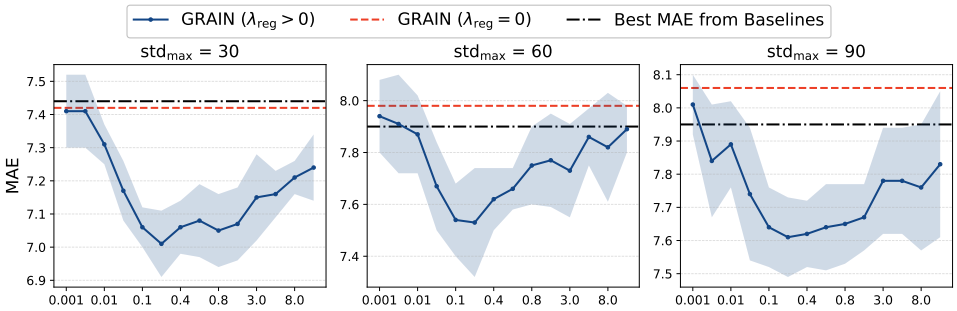


Figure 4: Hyperparameter sensitivity on λ_{reg}.

Additional Analyses. We further investigate the residual head behavior, the dynamics of the adaptive threshold τ, and the choice of discrepancy measure, with results provided in Appendix B.3. First, Figure B.7 shows that the learned residuals exhibit negligible correlation with the true labels (Pearson $R \approx 0.04$), whereas the regression predictions retain strong correlations, supporting the intended separation between residual correction and predictive signals. Second, Figure B.8 demonstrates that τ gradually decreases during training while maintaining larger values under higher noise levels, illustrating its adaptability to both noise severity and training dynamics. Finally, Figure B.6 compares UOT with balanced OT and cosine similarity as the discrepancy measure d in (11) and (12).

6 Conclusion

In this work, we introduce GRAIN, a dual-granularity framework for robust regression under noisy labels. At the fine-grained level, instance-specific auxiliary variables derived from a robust loss formulation decouple label noise from predictive signals. At the coarse-grained level, a binning-based contrastive regularization enforces an order-preserving feature geometry aligned with label proximity. The two levels are coupled via a histogram-based bridging mechanism that enables bidirectional information flow between structural and residual learning. Theoretical analysis and extensive experiments on synthetic and real-world datasets demonstrate that GRAIN consistently outperforms baselines across diverse settings. Future work includes extending the theoretical guarantees to weakly dependent settings and more general label noise generation models, and developing a formal analysis of the contrastive regularization and its interaction with residual learning. A detailed discussion of limitations, future directions, and broader impact is provided in Appendix C.

Acknowledgements

Yi is a Tier 1 Canada Research Chair in Data Science. Her research is supported by the Canada Research Chairs Program and the Natural Sciences and Engineering Research Council of Canada (NSERC), with partial support from the Canadian Statistical Sciences Institute (CANSSI) through its Collaborative Research Team program. Wang's research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grants program.

References

[1] Grace Y. Yi. *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer, New York, 2017.

[2] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, volume 31, 2018.

[3] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2021.

[4] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, pages 5907–5915. PMLR, 2019.

[5] Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *International Conference on Learning Representations*, 2020.

[6] Huaxiu Yao, Yiping Wang, Linjun Zhang, James Y Zou, and Chelsea Finn. C-mixup: Improving generalization in regression. In *Advances in neural information processing systems*, volume 35, pages 3361–3376, 2022.

[7] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud. Superloss: A generic loss for robust curriculum learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 4308–4319, 2020.

[8] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *International Conference on Learning Representations*, 2021.

[9] Hui Guo, Boyu Wang, and Grace Y. Yi. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. In *Advances in Neural Information Processing Systems*, 2023.

[10] Hui Guo, Grace Y. Yi, and Boyu Wang. Learning from noisy labels via conditional distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2024.

[11] Hui Guo, Grace Y. Yi, and Boyu Wang. Addressing both variable selection and misclassified responses with parametric and semiparametric methods. *Bernoulli*, 32(2):1303–1327, 2026.

[12] Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011.

[13] Chris Dongjoo Kim, Sangwoo Moon, Jihwan Moon, Dongyeon Woo, and Gunhee Kim. Sample selection via contrastive fragmentation for noisy label regression. *Advances in Neural Information Processing Systems*, 37:127561–127609, 2024.

[14] Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep regression with ordinal entropy. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Lawrence Stewart, Francis Bach, Quentin Berthet, and Jean-Philippe Vert. Regression as classification: Influence of task formulation on neural network features. In *International Conference on Artificial Intelligence and Statistics*, pages 11563–11582. PMLR, 2023.

[16] Gezheng Xu, Qi Chen, Qiuhao Zeng, Charles Ling, and Boyu Wang. Discretized density-guided source-free domain adaptation for regression. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

[17] Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

[18] Hicham Janati. *Advances in Optimal transport and applications to neuroscience*. PhD thesis, Institut Polytechnique de Paris, 2021.

[19] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[20] Albert Clapés, Ozan Bilici, Dariia Temirova, Egils Avots, Gholamreza Anbarjafari, and Sergio Escalera. From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2373–2382, 2018.

[21] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International journal of computer vision*, 101(3):437–458, 2013.

[22] Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019.

[23] Ewa Pietka, Arkadiusz Gertych, Sylwia Pospiech, Fei Cao, HK Huang, and Vicente Gilsanz. Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal roi extraction. *IEEE transactions on medical imaging*, 20(8):715–729, 2001.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[26] Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Annals of statistics*, 46(5):1904, 2018.

[27] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[28] Vladimir Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, pages 2877–2909, 2012.

[29] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

[30] Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471, 2023.

[31] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[32] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[33] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[34] Grace Y. Yi, Aurore Delaigle, and Paul Gustafson, editors. *Handbook of Measurement Error Models*. Chapman and Hall/CRC, 2021.

[35] Nancy Reid, Cristiano Varin, and Grace Y. Yi. *Likelihood and Its Extensions*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 2026.

[36] Pingbo Hu. *Statistical Learning of Noisy Data: Classification and Causal Inference with Measurement Error and Missingness*. PhD thesis, The University of Western Ontario, 2024.

[37] Jingyu Cui. *Multivariate Regression Analysis for Data with Measurement Error, Missing Values, and/or Sparsity Structures*. PhD thesis, The University of Western Ontario, 2023.

[38] Junhan Fang and Grace Y. Yi. Regularized matrix-variate logistic regression with response subject to misclassification. *Journal of Statistical Planning and Inference*, 217:106–121, 2022.

[39] Grace Y. Yi, Yanyuan Ma, Donna Spiegelman, and Raymond J. Carroll. Functional and structural methods with mixed measurement error and misclassification in covariates. *Journal of the American Statistical Association*, 110(510):681–696, 2015.

[40] Grace Y. Yi, Ying Yan, Xiaomei Liao, and Donna Spiegelman. Parametric regression analysis with covariate misclassification in main study/validation study designs. *The International Journal of Biostatistics*, 15(1):20170002, 2019.

[41] Jingyu Cui and Grace Y. Yi. Multivariate regression with measurement error: bias analysis and estimation. *Journal of Nonparametric Statistics*, 37(4):1257–1297, 2025.

[42] Jingyu Cui and Grace Y. Yi. Variable selection in multivariate regression models with measurement error in covariates. *Journal of Multivariate Analysis*, 202:105299, 2024.

[43] Grace Y. Yi, Xianming Tan, and Runze Li. Variable selection and inference procedures for marginal analysis of longitudinal data with missing observations and covariate measurement error. *The Canadian Journal of Statistics*, 43(4):498–518, 2015.

[44] Li-Pang Chen and Grace Y. Yi. Analysis of noisy survival data with graphical proportional hazards measurement error models. *Biometrics*, 77(3):956–969, 2021.

[45] Baojiang Chen, Ao Yuan, and Grace Y. Yi. Variable selection for proportional hazards models with high-dimensional covariates subject to measurement error. *The Canadian Journal of Statistics*, 49(2):397–420, 2021.

[46] Li-Pang Chen and Grace Y. Yi. Model selection and model averaging for analysis of truncated and censored data with measurement error. *Electronic Journal of Statistics*, 14(2):4054–4109, 2020.

[47] Pingbo Hu and Grace Y. Yi. Addressing label noise in causation classification via kernel embeddings. *Statistica Sinica*, 37(2), 2027.

[48] Yuan Bian, Grace Y. Yi, and Wenqing He. Empirical investigations of boosting with pseudo-outcome imputation for missing responses. *Statistics and Computing*, 35(2):40, 2025.

[49] Yuan Bian, Grace Y. Yi, and Wenqing He. A unified framework of analyzing missing data and variable selection using regularized likelihood. *Computational Statistics & Data Analysis*, 194:107919, 2024.

[50] Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A. Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *International Conference on Learning Representations*, 2023.

[51] Gezheng Xu, Li Yi, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Changjian Shui, A. Ian McLeod, Boyu Wang, and Charles Ling. Unraveling the mysteries of label noise in source-free domain adaptation: Theory and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):4007–4023, 2025.

[52] Gezheng Xu, Hui Guo, Li Yi, Charles Ling, Boyu Wang, and Grace Y. Yi. Revisiting source-free domain adaptation: a new perspective via uncertainty control. In *International Conference on Learning Representations*, 2025.

[53] Ruiyi Fang, Bingheng Li, Jingyu Zhao, Ruizhi Pu, Qiuhao Zeng, Gezheng Xu, Charles Ling, and Boyu Wang. Homophily enhanced graph domain adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[54] Ruiyi Fang, Jingyu Zhao, Shuo Wang, Ruizhi Pu, Bingheng Li, Jiale Cai, Zhihao Li, Zihao Jing, Jian Zhu, Song Tang, Charles Ling, and Boyu Wang. SAGA: Structural aggregation guided alignment with dynamic view and neighborhood order selection for multiview graph domain adaptation. In *International Conference on Learning Representations*, 2026.

[55] Ruiyi Fang, Changhong Chen, Qiang Wang, Charles Ling, Boyu Wang, and Hong Zhang. Transfer learning with graph neural networks to predict polymer solubility parameters. *npj Computational Materials*, 12(1):256, 2026.

[56] Yuan Bian, Yu Shi, Hui Guo, Grace Y. Yi, and Wenqing He. Physician effects in critical care: A causal inference approach through propensity weighting with parametric and super learning methods. *Journal of Data Science*, 23(1):130–148, 2025.

[57] Grace Y. Yi and Li-Pang Chen. Estimation of the average treatment effect with variable selection and measurement error simultaneously addressed for potential confounders. *Statistical Methods in Medical Research*, 32(4):691–711, 2023.

[58] Pingbo Hu and Grace Y. Yi. Inference about separable causal effects with longitudinal bivariate ordinal responses with missingness and censoring. *Statistics in Medicine*, 45(3–5):e70320, 2026.

[59] Yuan Bian, Xingche Guo, and Yuanjia Wang. Joint modeling for learning decision-making dynamics in behavioral experiments. *The Annals of Applied Statistics*, 19(4):3372–3393, 2025.

[60] Zhihao Li, Jiale Cai, Gezheng Xu, Hao Zheng, Qiuyue Li, Fan Zhou, Shichun Yang, Charles Ling, and Boyu Wang. Versatile transferable unlearnable example generator. In *Advances in Neural Information Processing Systems*, 2025.

[61] Zhihao Li, Gezheng Xu, Jiale Cai, Ruiyi Fang, Di Wu, Qicheng Lao, Charles Ling, and Boyu Wang. When priors backfire: On the vulnerability of unlearnable examples to pretraining. In *International Conference on Learning Representations*, 2026.

[62] Jiale Cai, Gezheng Xu, Zhihao Li, Ruiyi Fang, Ruizhi Pu, Di Wu, Qicheng Lao, Charles Ling, and Boyu Wang. FUSE: Full-spectrum unlearnable examples via spectral equalization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

[63] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

[64] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[65] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[66] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pages 5628–5637. PMLR, 2019.

Appendix: Theoretical Derivations, Implementation Details, and Additional Experiments

A Technical Details

Using [26], [27], and [28], we establish Theorem 4.2 under suitable regularity conditions. We first present additional notations and preliminary results, then we identify regularity conditions, and finally, we prove Theorem 4.2.

Notations. Let $\mathbb{R}_{\geq 0}$ denote the set of all nonnegative real values, and let $\mathbb{1}(\cdot)$ represent the indicator function. For any positive integer K, we denote $[K] \triangleq \{1, \ldots, K\}$. For any set $\Omega \subseteq \mathbb{R}^m$, where m is a positive integer, let $\mathcal{P}(\Omega)$ denote the set of all Borel probability measures on Ω, and let $\mathcal{M}_q^+(\Omega)$ denote the set of finite nonnegative Borel measures on Ω with finite q-th moments for $q \geq 1$. We further define $\mathcal{P}_q(\Omega) = \mathcal{P}(\Omega) \cap \mathcal{M}_q^+(\Omega)$ as the set of probability measures with finite q-th moments. For two vectors $a, a' \in \mathbb{R}^m$, we use $\langle a, a' \rangle$ to denote their dot product. For vector $a = (a_1, ..., a_p)^\top$ and $q \in [1, +\infty]$, the L_q norm of a is defined as $\|a\|_q = (\sum_{j=1}^p |a_j|^q)^{1/q}$ if $1 \leq q < \infty$, and $\|a\|_\infty = \max_j |a_j|$ if $q = +\infty$. For any norm $\|\cdot\|$ on $\mathbb{R}^m$, we use $\|\cdot\|_*$ to denote its dual norm, defined as $\|a\|_* = \sup\{\langle a, a' \rangle : \|a'\| \leq 1\}$. For a proper, closed, convex function $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, its convex conjugate (Fenchel conjugate) is defined as $f^*(s) = \sup_{a \in \mathbb{R}^m} \{\langle s, a \rangle - f(a)\}$. The Lipschitz modulus of a function $\hbar : \Omega \to \mathbb{R}$ is denoted by $\mathrm{Lip}(\hbar) = \sup_{a \neq a'} \frac{|\hbar(a) - \hbar(a')|}{\|a - a'\|}$.

Let $\mathbb{S}^{m-1} \triangleq \{u \in \mathbb{R}^m : \|u\|_2 = 1\}$, where $\|u\|_2$ denotes the L_2 norm for the vector u. For an $m \times m$ matrix V, we use $\|V\|_2$ to denote its spectral norm, defined as $\|V\|_2 = \sup_{u \in \mathbb{S}^{m-1}} \|Vu\|_2$. For V having real eigenvalues, let $\lambda_1(V), \ldots, \lambda_m(V)$ denote its eigenvalues with $\lambda_1(V) \geq \ldots \geq \lambda_m(V)$, and use $\lambda_{\min}(V)$ and $\lambda_{\max}(V)$ to stress the smallest and largest eigenvalues of V, respectively. We write $V \succ 0$ (or $V \succeq 0$) if V is a positive (or semipositive) definite matrix.

A.1 Preliminaries

Definition A.1 (OT, Kantorovitch formulation, 29)**.** Let Ω denote a compact metric space equipped with a non-negative cost function $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$. For any $P_1, P_2 \in \mathcal{P}(\Omega)$, the *optimal transport cost* between P_1 and P_2 is defined as

$$\mathrm{OT}(P_1, P_2) \triangleq \inf_{\Pi \in \mathrm{Cpl}(P_1, P_2)} \int_{\Omega \times \Omega} \boldsymbol{c}(u_1, u_2) d\Pi(u_1, u_2), \tag{A.1}$$

where $\mathrm{Cpl}(P_1, P_2)$, commonly referred to as *transport plans*, denotes the set of couplings between P_1 and P_2, i.e., probability measures on $\Omega \times \Omega$ with marginals P_1 and P_2. Any minimizer $\Pi^* \in \mathrm{Cpl}(P_1, P_2)$ of (A.1) is called an *optimal transport plan* between P_1 and P_2.

Definition A.2 (UOT, 30)**.** Consider the setup in Definition A.1. For any $P_1, P_2 \in \mathcal{P}(\Omega)$ and $\gamma, \gamma' > 0$, the *entropy-regularized unbalanced optimal transport cost* between P_1 and P_2 is defined as

$$\begin{aligned} \mathrm{UOT}_\gamma^{\gamma'}(P_1, P_2) \triangleq \inf_{\Pi \in \mathcal{M}_2^+(\Omega \times \Omega)} & \int_{\Omega \times \Omega} \boldsymbol{c}(u_1, u_2) d\Pi(u_1, u_2) \\ & + \gamma \mathrm{KL}(\Pi_1 \| P_1) + \gamma \mathrm{KL}(\Pi_2 \| P_2) \\ & + \gamma' \mathrm{KL}(\Pi \| P_1 \otimes P_2), \end{aligned}$$

where Π_1 and Π_2 denote the first and second marginal distributions of Π, respectively, $\mathrm{KL}(\cdot \| \cdot)$ is the Kullback–Leibler (KL) divergence, and $P_1 \otimes P_2$ is the product measure of P_1 and P_2. When $\gamma' = 0$, we write the UOT cost as $\mathrm{UOT}_\gamma(P_1, P_2)$.

Remark A.3. Unlike balanced OT in Definition A.1, which enforces exact marginal matching and is therefore sensitive to outliers, UOT relaxes these hard constraints into a soft alignment, penalizing large deviations while permitting controlled mass mismatch. This flexibility allows the transport plan to deviate from the original marginals, effectively downweighting outliers and improving robustness.

Lemma A.4 ([26])**.** *For any positive integer $d \geq 1$ and convex loss function $\ell : \mathbb{R}^d \to \mathbb{R}$, let $\mathtt{D}_\ell(\beta_1, \beta_2) \triangleq \ell(\beta_1) - \ell(\beta_2) - \langle \nabla \ell(\beta_2), \beta_1 - \beta_2 \rangle$ and $\bar{\mathtt{D}}_\ell(\beta_1, \beta_2) \triangleq \mathtt{D}_\ell(\beta_1, \beta_2) + \mathtt{D}_\ell(\beta_2, \beta_1)$ for $\beta_1, \beta_2 \in \mathbb{R}^d$. Then, for any $\beta_\eta \triangleq \beta_1 + \eta(\beta_2 - \beta_1)$ with $\eta \in [0, 1]$, $\bar{\mathtt{D}}_\ell(\beta_1, \beta_\eta) \leq \eta \bar{\mathtt{D}}_\ell(\beta_1, \beta_2)$.*

Lemma A.5. *Let $A, B \in \mathbb{R}^{m_1 \times m_2}$, with m_1 and m_2 denoting two positive integers. Then the following results hold.*

(i) $\|A\|_2 = \sup_{u \in \mathbb{S}^{m_2-1}} \|Au\|_2 = \sup_{u \in \mathbb{S}^{m_1-1}, v \in \mathbb{S}^{m_2-1}} u^\top A v$; $\|A\|_2^2 = \|A^\top A\|_2 = \lambda_{\max}(A^\top A)$. *Specifically, for a vector $a \in \mathbb{R}^{m_1}$, we have $\|a\|_2 = \sup_{u \in \mathbb{S}^{m_1-1}} u^\top a$.*

(ii) If $m_1 = m_2 = m$ and A is symmetric, then $\sup_{u \in \mathbb{S}^{m-1}} u^\top A u = \lambda_{\max}(A)$ and $\inf_{u \in \mathbb{S}^{m-1}} u^\top A u = \lambda_{\min}(A)$. Additionally, $\|A\|_2 = \max_{1 \le i \le m} |\lambda_i(A)| = \max\{|\lambda_{\max}(A)|, |\lambda_{\min}(A)|\}$.

(iii) (Weyl's inequality) If $m_1 = m_2 = m$ and A and B are symmetric, then

- *$\lambda_{i+j-1}(A+B) \le \lambda_i(A) + \lambda_j(B)$ with $1 \le i, j, i+j-1 \le m$;*
- *$\lambda_i(A) + \lambda_j(B) \le \lambda_{i+j-m}(A+B)$ with $1 \le i, j, i+j-m \le m$.*

In particular, $\lambda_{\min}(A+B) \ge \lambda_{\min}(A) + \lambda_{\min}(B)$, and $\lambda_{\max}(A+B) \le \lambda_{\max}(A) + \lambda_{\max}(B)$.

Lemma A.6 (Modified from Corollaries 1.12 and 1.13 in [28])**.** *Let ζ denote a random vector in $\mathbb{R}^m$. Assume that there exist matrix $V_0 \in \mathbb{R}^{m \times m}$ and constant $\mathtt{g} > 0$ such that*

$$\sup_{u \in \mathbb{R}^m} \log \mathbb{E} \left\{ \exp\left(t \frac{u^\top \zeta}{\|V_0 u\|_2} \right) \right\} \le t^2/2, \ \ |t| \le \mathtt{g}. \tag{A.2}$$

Let $D_0 \in \mathbb{R}^{m \times m}$ denote a given positive definite matrix, and define $\mathbb{B} \triangleq D_0^{-1/2} V_0 D_0^{-1/2}$. Let $\mathtt{m} \triangleq tr(\mathbb{B}^2)$ and $\mathtt{v} \triangleq \sqrt{2tr(\mathbb{B}^4)}$, and assume $\mathtt{g} \ge \sqrt{2\mathtt{m}}$. Denote $\mathtt{c}_1 \triangleq \mathtt{g}^2/(\mathtt{m}+\mathtt{g}^2)$, $\mathtt{c}_2 \triangleq \sqrt{\mathtt{g}^2/\mathtt{c}_1^2 - \mathtt{m}/\mathtt{c}_1} = \sqrt{\mathtt{m}+\mathtt{g}^2}$, $\mathtt{c}_3 \triangleq \mathtt{c}_1\mathtt{c}_2 = \sqrt{\mathtt{g}^2 - \mathtt{c}_1\mathtt{m}} = \mathtt{g}^2/\sqrt{\mathtt{m}+\mathtt{g}^2}$, and $\mathtt{c}_4 \triangleq \frac{1}{2}\left\{\mathtt{c}_3\mathtt{c}_2 + \log\det(I_m - \mathtt{c}_1\mathbb{B}^2)\right\}$. Then

(a) for any $\mathtt{x} \le \mathtt{c}_4$,

$$\mathbb{P}\left\{\|D_0^{-1}\zeta\|_2^2 \ge \hbar_1(\mathtt{x}, \mathbb{B})\right\} \le 2e^{-\mathtt{x}} + 8.4e^{-\mathtt{c}_4},$$

where

$$\hbar_1(\mathtt{x}, \mathbb{B}) \triangleq \begin{cases} \mathtt{m} + \sqrt{2\mathtt{x}\mathtt{v}}, & \mathtt{x} \le \mathtt{v}/18, \\ \mathtt{m} + 6\mathtt{x}, & \mathtt{v}/18 < \mathtt{x} \le \mathtt{c}_4. \end{cases}$$

(b) for any $\mathtt{x} > \mathtt{c}_4$,

$$\mathbb{P}\left\{\|D_0^{-1}\zeta\|_2^2 \ge \hbar_2(\mathtt{x}, \mathbb{B})\right\} \le 8.4e^{-\mathtt{x}},$$

where $\hbar_2(\mathtt{x}, \mathbb{B}) \triangleq |\mathtt{c}_2 + 2(\mathtt{x} - \mathtt{c}_4)/\mathtt{c}_3|^2$.

Lemma A.7 (Sub-gaussian properties; [31], Proposition 2.5.2)**.** *For $j = 1, \ldots, 5$, let K_j denote positive constants that may differ from each other by at most an absolute constant factor. Let ζ be a random variable. Then the following statements are equivalent:*

(i) Tail probability: $\mathbb{P}(|\zeta| \ge t) \le 2\exp(-t^2/K_1^2)$ for all $t \ge 0$.

(ii) Moment bound: $\|\zeta\|_{L_p} \triangleq \{\mathbb{E}(|\zeta|^p)\}^{1/p} \le K_2\sqrt{p}$ for all $p \ge 1$.

(iii) Bound for the moment generating function: $\mathbb{E}\{\exp(t^2\zeta^2)\} \le \exp(K_3^2 t^2)$ for all t with $|t| \le \frac{1}{K_3}$.

(iv) Bound for a special moment generating function: $\mathbb{E}\{\exp(\zeta^2/K_4^2)\} \le 2$.

Moreover, if $\mathbb{E}\zeta = 0$ then properties (i)–(iv) are also equivalent to the following one:

(v) Bound for the moment generating function: $\mathbb{E}\{\exp(t\zeta)\} \le \exp(K_5^2 t^2)$ for all $t \in \mathbb{R}$.

Definition A.8 (Sub-Gaussian random variables and vectors)**.**

(i) [[31], Definition 2.5.6] A random variable ζ satisfying one of the statements from (i) to (iv) in Lemma A.7 is called a *sub-gaussian random variable*. The *sub-gaussian norm* of ζ, denoted $\|\zeta\|_{\psi_2}$, is defined to be the smallest possible K_4 in statement (iv), i.e.,

$$\|\zeta\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\exp(\zeta^2/t^2) \le 2\right\}.$$

Here, the notation ψ_2 comes from the Orlicz function $\psi_\alpha(x) = \exp(x^\alpha) - 1$ with $\alpha = 2$, which is commonly used to characterize sub-Gaussian tail behavior.

(ii) [[31], Definition 3.4.1] A random vector ζ in $\mathbb{R}^m$ is called *sub-Gaussian* if for any $u \in \mathbb{R}^m$, $\langle \zeta, u\rangle$ is a sub-Gaussian random variable. The *sub-Gaussian norm* of ζ is defined as

$$\|\zeta\|_{\psi_2} = \sup_{u\in\mathbb{S}^{m-1}} \|\langle \zeta, u\rangle\|_{\psi_2}.$$

Lemma A.9 ([31], (4.22); [32], Remark 5.40)**.** *Assume that A is an $m \times d$ matrix whose rows A_j are independent sub-gaussian random vectors in $\mathbb{R}^d$, with second moment matrix $\Sigma_A = \mathbb{E}(A_j A_j^\top)$ for $j \in [m]$. Then for every $w \ge 0$, with probability at least $1 - 2e^{-C_1 w^2}$,*

$$\left\| \frac{1}{m} A^\top A - \Sigma_A \right\|_2 \le \max(\epsilon, \epsilon^2), \text{ with } \epsilon = C_2\sqrt{\frac{d}{m}} + \frac{w}{\sqrt{m}},$$

where C_1 and C_2 are positive constants depending only on the subgaussian norms of the rows.

Lemma A.10 (Conjugacy of infimal convolution)**.** *Let f and g be proper, lower semicontinuous, convex functions on $\mathbb{R}$. Their infimal convolution is defined by*

$$(f\square g)(x) = \inf_{x'\in\mathbb{R}} \{f(x') + g(x - x')\}.$$

Then the following hold:

(i) ***(Dual identity)*** *For any proper convex f and g such that $f\square g$ is proper,*

$$(f\square g)^* = f^* + g^*.$$

(ii) ***(Primal recovery)*** *If in addition $f\square g$ is lower semicontinuous, for instance if $\operatorname{int}(\operatorname{dom} f) \cap \operatorname{dom} g \ne \emptyset$, then by the biconjugate theorem,*

$$f\square g = (f^* + g^*)^*.$$

(iii) ***(Converse)*** *If h is proper, lower semicontinuous, convex and $h^* = f^* + g^*$, then $h = f\square g$.*

Lemma A.11 (Smoothness–strong convexity duality)**.** *Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous, convex function. Then f has L-Lipschitz gradient if and only if f^* is $\frac{1}{L}$-strongly convex, i.e.,*

$$f^*(tw + (1-t)w') \le tf^*(w) + (1-t)f^*(w') - \frac{1}{2L}t(1-t)(w-w')^2, \quad \forall\, w, w' \in \mathbb{R},\ t \in [0,1].$$

Equivalently, the function $w \mapsto f^*(w) - \frac{1}{2L}w^2$ is convex.

Lemma A.12 (Quadratic infimal convolution criterion)**.** *Let $q(x) = x^2/2$, and let ρ be a proper, lower semicontinuous, convex, and differentiable function whose derivative is 1-Lipschitz. Then*

$$\psi(w) := \rho^*(w) - q^*(w) = \rho^*(w) - \frac{1}{2}w^2$$

is convex. Consequently, defining $\phi := \psi^*$, we have

$$\rho(\nu) = \inf_{\xi\in\mathbb{R}} \left\{ \frac{1}{2}\xi^2 + \phi(\nu - \xi) \right\}.$$

Moreover, ϕ is uniquely determined in the class of proper lower semicontinuous convex functions.

Proof. Since ρ is proper, lower semicontinuous, convex, and has 1-Lipschitz derivative, Lemma A.11 (with $L = 1$) implies that ρ^* is 1-strongly convex, i.e., the function $w \mapsto \rho^*(w) - \frac{1}{2}w^2$ is convex. Since $q^*(w) = \frac{1}{2}w^2$, this shows that $\psi = \rho^* - q^*$ is convex.

As ρ^* is proper and lower semicontinuous, and q^* is continuous and finite-valued, it follows that ψ is also proper and lower semicontinuous. Define $\phi := \psi^*$. Then ϕ is proper, lower semicontinuous, and convex. By the Fenchel–Moreau theorem,

$$\phi^* = \psi^{**} = \psi = \rho^* - q^*.$$

Hence,

$$q^* + \phi^* = \rho^*.$$

Applying Lemma A.10(iii) yields

$$\rho = q \square \phi,$$

which gives the desired representation.

Finally, since $\phi^* = \rho^* - q^*$ is uniquely determined by ρ, the function ϕ is uniquely determined (as a proper lower semicontinuous convex function) via the conjugacy relation. □

Lemma A.13. *Under Assumption 4.1, $\ell_\tau(\cdot)$ defined in (3) is convex.*

Proof. Since $\ell_\tau(\cdot)$ is an even function, its derivative $\ell'_\tau(\cdot)$ is odd. It therefore suffices to show that ℓ'_τ is non-decreasing on $[0, \infty)$ and that $\ell'_\tau(0) = 0$, as these together imply ℓ'_τ is non-decreasing on all of $\mathbb{R}$.

For $\nu \geq 0$,

$$\ell'_\tau(\nu) = \begin{cases} \nu, & 0 \leq \nu \leq \tau, \\ \dfrac{\tau}{\ell'(\tau)}\ell'(\nu), & \nu > \tau. \end{cases}$$

We have $\ell'_\tau(0) = 0$, and the derivative is continuous at $\nu = \tau$ since

$$\frac{\tau}{\ell'(\tau)}\ell'(\tau) = \tau.$$

On $[0, \tau)$, $\ell'_\tau(\nu) = \nu$ is non-decreasing. On (τ, ∞), since ℓ is convex by Assumption 4.1, ℓ' is non-decreasing, and hence $\ell'_\tau(\nu) = \frac{\tau}{\ell'(\tau)}\ell'(\nu)$ is also non-decreasing. Continuity at τ then ensures ℓ'_τ is non-decreasing on all of $[0, \infty)$, and therefore on $\mathbb{R}$. Hence $\ell_\tau(\cdot)$ is convex. □

A.2 Regularity Conditions

(i) The zero-mean error variables ε' and ε'' in models (1) and (2) are independent and also independent of X. Without loss of generality, we assume $\mathbb{E}(\varepsilon'^2) = \mathbb{E}(\varepsilon''^2) = 1$.

(ii) Let $\bar{\varsigma}^2(\mathrm{X}) \triangleq 1 + \mathbb{E}_{\varepsilon'}\left\{\varsigma^2(\mathrm{X}, \langle\varphi(\mathrm{X}), \theta^*_{\text{reg}}\rangle + \varepsilon')|\mathrm{X}\right\}$. Assume that $\alpha_1 \triangleq \mathbb{E}\{\bar{\varsigma}^{2+\iota_1}(\mathrm{X})\} < \infty$ for some $\iota_1 > 0$ and denote $\sigma^2 \triangleq \mathbb{E}\{\bar{\varsigma}^2(\mathrm{X})\}$.

(iii) For $\mathrm{Z} = \varphi(\mathrm{X})$ defined in Section 4, let $\Sigma \triangleq \mathbb{E}(\mathrm{Z}\mathrm{Z}^\top)$ and $\nu^* \triangleq \varepsilon' + \varsigma(\mathrm{X}, \mathrm{Y})\varepsilon''$. Assume that Σ is strictly positive definite, and let $\zeta \triangleq \Sigma^{-1/2}\mathrm{Z}$. We assume the following conditions.

  (a) ζ is a sub-Gaussian random vector. Denote by $K_j^{(\zeta)}$, $1 \leq j \leq 4$, the constants associated with ζ in Lemma A.7 (i)–(iv), and let $\alpha_2 \triangleq \min_{1\leq j\leq 4} K_j^{(\zeta)}$.

  (b) There exists a constant $\kappa \geq 0$ such that $|\nu^*|^\kappa\zeta$ is a sub-Gaussian random vector. Denote by $K_j^{(\kappa)}$, $1 \leq j \leq 4$, the constants associated with $|\nu^*|^\kappa\zeta$ in Lemma A.7 (i)–(iv), and let $\alpha_3 \triangleq \min_{1\leq j\leq 4} K_j^{(\kappa)}$.

A.3 Distributionally Robust Extension

In this subsection, we examine a distributionally robust optimization (DRO) perspective on learning with noisy labels. Instead of minimizing the empirical risk $\Re^{\tau}_{\text{reg}}(\theta_{\text{reg}})$ under the empirical distribution $\widehat{P}^{\dagger}$ defined in Section 4.1, we consider the worst-case risk over a neighborhood of distributions:

$$\Re^{\tau,\rho}_{\text{reg}}(\theta_{\text{reg}}) = \sup_{Q\in\Gamma_\rho(\widehat{P}^{\dagger})} \mathbb{E}_{(\mathrm{Z},\widetilde{\mathrm{Y}})\sim Q}\left\{\ell_\tau(\langle \mathrm{Z},\theta_{\text{reg}}\rangle - \widetilde{\mathrm{Y}})\right\}, \tag{A.3}$$

where $\ell_\tau(\cdot)$ is defined in (3), and

$$\Gamma_\rho(\widehat{P}^{\dagger}) = \left\{Q\in\mathcal{P}(\mathcal{Z}\times\mathcal{Y}) : \mathscr{d}(Q,\widehat{P}^{\dagger})\le\rho\right\}, \tag{A.4}$$

denotes an *ambiguity set* centered at $\widehat{P}^{\dagger}$, where $\mathscr{d}$ is a discrepancy metric between probability measures and $\rho>0$ controls the size of the uncertainty set. This formulation captures distributional perturbations arising from label noise.

We construct $\Gamma_\rho(\widehat{P}^{\dagger})$ using an optimal transport (OT) discrepancy, which is given in Definition A.1. Specifically, let the cost function $\mathscr{c}:\Omega\times\Omega\to\mathbb{R}_+$ be defined as $\mathscr{c}(u_1,u_2)=\|u_1-u_2\|$ for a norm $\|\cdot\|$ on $\Omega=\mathcal{Z}\times\mathcal{Y}$. Under this choice, the DRO problem in (A.3) admits the following dual reformulation.

Lemma A.14 (33)**.** *Assume that $\ell_\tau(\cdot)$ is convex and L-Lipschitz continuous. Then, $\inf_{\theta_{reg}}\Re^{\tau,\rho}_{reg}(\theta_{reg})$ is equivalent to the following optimization problem:*

$$\inf_{\theta_{\text{reg}}}\left[\mathbb{E}_{(\mathrm{Z},\widetilde{\mathrm{Y}})\sim\widehat{P}^{\dagger}}\left\{\ell_\tau\left(\langle \mathrm{Z},\theta_{reg}\rangle-\widetilde{\mathrm{Y}}\right)\right\}+\rho Lip(\ell_\tau)\|(\theta_{reg}^{\top},-1)^{\top}\|_*\right],$$

where Lip(ℓ_τ) denotes the Lipschitz modulus of the loss function ℓ_τ, and $\|\cdot\|_*$ represents the dual norm of $\|\cdot\|$.

Lemma A.14 establishes an exact regularized reformulation of the DRO problem (A.3) under its stated conditions. In particular, when $\ell(t)=t$, the piecewise loss ℓ_τ reduces to the Huber loss with threshold τ, which is globally Lipschitz continuous with $\mathrm{Lip}(\ell_\tau)=\tau$. For more general choices of ℓ, however, ℓ_τ need not be globally Lipschitz, and the equivalence in Lemma A.14 does not necessarily apply. For example, if the norm $\|\cdot\|$ in the cost function $\mathscr{c}$ is set to the L_q norm for any $1<q<\infty$, then the corresponding dual norm $\|\cdot\|_*$ is the $L_{q'}$ norm with $1/q+1/q'=1$. In particular, when $q=2$, both norms $\|\cdot\|$ and $\|\cdot\|_*$ are identical to the L_2 norm.

Motivated by the connection between DRO and regularization, we consider the following ridge-regularized empirical objective:

$$\Re^{\tau,\lambda}_{\text{reg}}(\theta_{\text{reg}}) \triangleq \frac{1}{N}\sum_{i=1}^{N}\ell_\tau(\langle \mathrm{Z}_i,\theta_{\text{reg}}\rangle-\widetilde{\mathrm{Y}}_i)+\frac{\lambda}{2}\|\theta_{\text{reg}}\|_2^2, \tag{A.5}$$

where $\lambda\ge 0$ is a regularization parameter. Unlike the exact dual-norm regularization in Lemma A.14, the squared L_2 penalty in (A.5) is adopted as a DRO-motivated regularization choice. Let $\widehat{\theta}^{\tau,\lambda}_{\text{reg}}$ denote a minimizer of (A.5). The following result establishes statistical guarantees for this ridge-regularized estimator under the general conditions of Theorem 4.2. When $\lambda=0$, it recovers Theorem 4.2.

Theorem A.15. *Assume that the conditions in Theorem 4.2 hold and that $\lambda\le\mathscr{C}_3\sqrt{\frac{d+s}{N}}$ for some constant $\mathscr{C}_3>0$. Then,*

- *(i) with probability at least $1-5e^{-s}$, $\|\widehat{\theta}^{\tau,\lambda}_{reg}-\theta^*_{reg}\|_\Sigma\le\mathscr{C}^{\dagger}\sqrt{\frac{d+s}{N}}$ for some constant $\mathscr{C}^{\dagger}>0$;*
- *(ii) with probability at least $1-7e^{-s}$, for any future input X_{new} with feature $\mathrm{Z}_{new}\triangleq\varphi(\mathrm{X}_{new})$, the associated prediction error satisfies: $err\triangleq\left|\langle \mathrm{Z}_{new},\widehat{\theta}^{\tau,\lambda}_{reg}\rangle-\langle \mathrm{Z}_{new},\theta^*_{reg}\rangle\right|\le\mathscr{C}^{\ddagger}\sqrt{\frac{ds+s^2}{N}}$ for some constant $\mathscr{C}^{\ddagger}>0$.*

A.4 Proof of Theorem A.15 (i) (Theorem 4.2 (i))

Since the DRO-induced empirical risk in (A.5) reduces to the empirical risk in (4) when $\lambda = 0$, the corresponding minimizer $\widehat{\theta}^{\tau,\lambda}\text{reg}$ reduces to $\widehat{\theta}^{\tau}\text{reg}$. Hence, it suffices to prove Theorem A.15 (i) only.

Recall that in Section 4.1, with a fixed feature extractor $\varphi(\cdot)$, the data are assumed to follow the generative mechanism specified by (1) and (2), together with the setup at the beginning of Section 4.1:

$$\mathrm{Y} = \langle \mathrm{Z}, \theta^*_{\text{reg}} \rangle + \varepsilon' \tag{A.6}$$

$$\widetilde{\mathrm{Y}} = \mathrm{Y} + \varsigma(\mathrm{X}, \mathrm{Y})\varepsilon'' \tag{A.7}$$

We consider the piecewise loss ℓ_τ in (3) and rewrite the DRO-induced empirical risk in (A.5) as:

$$\begin{aligned}\mathcal{L}(\theta_{\text{reg}}) \triangleq & \frac{1}{N}\sum_{i=1}^{N} \ell_\tau(|\nu_i|) + \frac{\lambda}{2}\|\theta_{\text{reg}}\|_2^2 \\ = & \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\nu_i^2 \mathbb{1}(|\nu_i| \le \tau) + \frac{\tau}{\ell'(\tau)}\left\{\ell(|\nu_i|) - \ell(\tau)\right\}\mathbb{1}(|\nu_i| > \tau)\right] + \frac{\lambda}{2}\|\theta_{\text{reg}}\|_2^2, \end{aligned} \tag{A.8}$$

where $\nu_i \triangleq \widetilde{Y}_i - \langle \mathrm{Z}_i, \theta_{\text{reg}} \rangle$. Additionally,

$$\nabla \mathcal{L}(\theta_{\text{reg}}) = -\frac{1}{N}\sum_{i=1}^{N} \text{sgn}(\nu_i)\ell'_\tau(|\nu_i|)\mathrm{Z}_i + \lambda\theta_{\text{reg}}; \tag{A.9}$$

$$\nabla^2 \mathcal{L}(\theta_{\text{reg}}) = \frac{1}{N}\sum_{i=1}^{N} \ell''_\tau(|\nu_i|)\mathrm{Z}_i\mathrm{Z}^\top + \lambda I_{d\times d}, \tag{A.10}$$

where $\text{sgn}(\cdot)$ denotes the sign function, defined as $\text{sgn}(x) = 1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$, and $\text{sgn}(0) = 0$; $I_{d\times d}$ representing the $d \times d$ identity matrix; and

$$\begin{aligned}\ell'_\tau(|\nu_i|) \triangleq & |\nu_i|\mathbb{1}(|\nu_i| \le \tau) + \frac{\tau\ell'(|\nu_i|)}{\ell'(\tau)}\mathbb{1}(|\nu_i| > \tau) \\ \le & |\nu_i|\mathbb{1}(|\nu_i| \le \tau) + \min(|\nu_i|, \frac{\tau|\nu_i|^\kappa}{\ell'(\tau)}); \end{aligned} \tag{A.11}$$

$$\ell''_\tau(|\nu_i|) \triangleq \mathbb{1}(|\nu_i| \le \tau) + \frac{\tau\ell''(|\nu_i|)}{\ell'(\tau)}\mathbb{1}(|\nu_i| > \tau), \tag{A.12}$$

where the inequality in (A.11) follows from $\frac{\tau\ell'(|\nu_i|)}{\ell'(\tau)} \le |\nu_i|$ for $|\nu_i| > \tau$, which is implied by the assumption in Theorem A.15 that $\frac{\ell'(t)}{t}$ is nonincreasing for $t \ge \tau$, together with the assumption $\ell'(t) \le t^\kappa$ for $t > \tau$ in Theorem A.15.

For $\eta \in (0,1)$, let $\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) \triangleq \theta^*_{\text{reg}} + \eta(\widehat{\theta}^{\tau,\lambda}_{\text{reg}} - \theta^*_{\text{reg}})$ denote an "intermediate" estimator of θ^*_{reg}. Then,

$$\begin{aligned}\left\langle \nabla\mathcal{L}(\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)) - \nabla\mathcal{L}(\theta^*_{\text{reg}}), \widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}} \right\rangle &\le \eta\left\langle \nabla\mathcal{L}(\widehat{\theta}^{\tau,\lambda}_{\text{reg}}) - \nabla\mathcal{L}(\theta^*_{\text{reg}}), \widehat{\theta}^{\tau,\lambda}_{\text{reg}} - \theta^*_{\text{reg}} \right\rangle \\ &= \eta\left\langle -\nabla\mathcal{L}(\theta^*_{\text{reg}}), \widehat{\theta}^{\tau,\lambda}_{\text{reg}} - \theta^*_{\text{reg}} \right\rangle \\ &= -\left\langle \nabla\mathcal{L}(\theta^*_{\text{reg}}), \widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}} \right\rangle, \end{aligned} \tag{A.13}$$

where the first step is due to Lemma A.4 and Lemma A.13, the second step holds since $\widehat{\theta}^{\tau,\lambda}_{\text{reg}}$ is a minimizer of $\mathcal{L}(\theta_{\text{reg}})$, and the last step comes from the definition of $\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)$.

Next, we examine the left-hand side of (A.13). By the mean value theorem, there exists a convex combination of $\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)$ and θ^*_{reg}, denoted by $\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)$, such that $\nabla\mathcal{L}(\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)) - \nabla\mathcal{L}(\theta^*_{\text{reg}}) = \nabla^2\mathcal{L}(\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta))\left\{\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}}\right\}$. Therefore, using $\Sigma \triangleq \mathbb{E}(\mathrm{Z}\mathrm{Z}^\top)$ and Lemma A.5 (ii), we obtain

$$\begin{aligned}&\left\langle \nabla\mathcal{L}(\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)) - \nabla\mathcal{L}(\theta^*_{\text{reg}}), \widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}} \right\rangle \\ = &\left\{\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}}\right\}^\top \nabla^2\mathcal{L}(\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta))\left\{\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}}\right\}\end{aligned}$$

$$= \left[\Sigma^{1/2}\left\{\widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\}\right]^\top \left\{\Sigma^{-1/2}\nabla^2\mathcal{L}(\widetilde{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta))\Sigma^{-1/2}\right\}\left[\Sigma^{1/2}\left\{\widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\}\right]$$
$$\geq \left\|\Sigma^{1/2}\left\{\widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\}\right\|_2^2 \lambda_{\min}\left(\Sigma^{-1/2}\nabla^2\mathcal{L}(\widetilde{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta))\Sigma^{-1/2}\right), \tag{A.14}$$

which identifies a lower bound of the right-hand side of (A.13).

On the other hand, using Σ and the Cauchy–Schwarz inequality, the right-hand side of (A.13) is upper bounded as:

$$\begin{aligned}
-\left\langle \nabla\mathcal{L}(\theta^*_{\mathrm{reg}}), \widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\rangle &= \left|\left\{\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\}^\top\left\{\widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\}\right| \\
&= \left|\left\{\Sigma^{-1/2}\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\}^\top \Sigma^{1/2}\left\{\widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\}\right| \\
&\leq \left\|\Sigma^{-1/2}\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\|_2 \left\|\Sigma^{1/2}\left\{\widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\}\right\|_2,
\end{aligned} \tag{A.15}$$

where the first step holds since $-\left\langle \nabla\mathcal{L}(\theta^*_{\mathrm{reg}}), \widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\rangle$ admits a positive lower bound. Specifically, it is lower bounded by (A.13), which in turn is strictly positive by (A.14).

Comparing (A.14) and (A.15) via (A.13), we obtain that

$$\left\|\Sigma^{1/2}\left\{\widehat{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta) - \theta^*_{\mathrm{reg}}\right\}\right\|_2 \lambda_{\min}\left(\Sigma^{-1/2}\nabla^2\mathcal{L}(\widetilde{\theta}^{\tau,\lambda}_{\mathrm{reg}}(\eta))\Sigma^{-1/2}\right) \leq \left\|\Sigma^{-1/2}\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\|_2. \tag{A.16}$$

Step 1: Upper bounding $\left\|\Sigma^{-1/2}\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\|_2$ in (A.16).

By the triangle inequality, $\left\|\Sigma^{-1/2}\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\|_2$ is upper bounded as:

$$\left\|\Sigma^{-1/2}\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\|_2 \leq \underbrace{\left\|\Sigma^{-1/2}\mathbb{E}\left\{\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\}\right\|_2}_{\text{(I)}} + \underbrace{\left\|\Sigma^{-1/2}\left[\nabla\mathcal{L}(\theta^*_{\mathrm{reg}}) - \mathbb{E}\left\{\nabla\mathcal{L}(\theta^*_{\mathrm{reg}})\right\}\right]\right\|_2}_{\text{(II)}}. \tag{A.17}$$

Step 1.1: Examining (I).

For $i \in [n]$, letting $\zeta_i \triangleq \Sigma^{-1/2}\mathrm{Z}_i$ and $\nu^*_i = \widetilde{Y}_i - \langle \mathrm{Z}_i, \theta^*_{\mathrm{reg}}\rangle$. Then by (A.6) and (A.7), $\nu^*_i = \widetilde{Y}_i - \langle \mathrm{Z}_i, \theta^*_{\mathrm{reg}}\rangle = \varepsilon'_i + \varsigma(\mathrm{X}_i, \mathrm{Y}_i)\varepsilon''_i$, yielding that $\mathbb{E}(\nu^*_i) = 0$ and $\mathbb{E}(\nu^*_i\zeta_i) = 0$. By (A.9), we have

$$\begin{aligned}
\text{(I)} &= \left\|-\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\{\mathrm{sgn}(\nu^*_i)\ell'_\tau(|\nu^*_i|)\Sigma^{-1/2}\mathrm{Z}_i\right\} + \lambda\Sigma^{-1/2}\theta^*_{\mathrm{reg}}\right\|_2 \\
&\leq \left\|-\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\{\mathrm{sgn}(\nu^*_i)\ell'_\tau(|\nu^*_i|)\Sigma^{-1/2}\mathrm{Z}_i\right\}\right\|_2 + \lambda\|\Sigma^{-1/2}\theta^*_{\mathrm{reg}}\|_2 \\
&= \sup_{u\in\mathbb{S}^{d-1}}\left[-\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\{\mathrm{sgn}(\nu^*_i)\ell'_\tau(|\nu^*_i|)u^\top\zeta_i\right\}\right] + \lambda\|\Sigma^{-1/2}\theta^*_{\mathrm{reg}}\|_2,
\end{aligned} \tag{A.18}$$

where the last step is due to Lemma A.5 (i).

For any $u \in \mathbb{S}^{d-1}$, we have that by (A.11),

$$\begin{aligned}
&\left|\mathbb{E}\left\{\mathrm{sgn}(\nu^*_i)\ell'_\tau(|\nu^*_i|)u^\top\zeta_i\right\}\right| \\
&= \left|\mathbb{E}\left[\mathrm{sgn}(\nu^*_i)\left\{|\nu^*_i|\mathbb{1}(|\nu^*_i| \leq \tau)u^\top\zeta_i + \frac{\tau\ell'(|\nu^*_i|)}{\ell'(\tau)}\mathbb{1}(|\nu^*_i| > \tau)u^\top\zeta_i\right\}\right]\right| \\
&= \left|\mathbb{E}\left[\mathrm{sgn}(\nu^*_i)\left\{|\nu^*_i|u^\top\zeta_i - |\nu^*_i|\mathbb{1}(|\nu^*_i| > \tau)u^\top\zeta_i + \frac{\tau\ell'(|\nu^*_i|)}{\ell'(\tau)}\mathbb{1}(|\nu^*_i| > \tau)u^\top\zeta_i\right\}\right]\right| \\
&= \left|\mathbb{E}\left[-\mathrm{sgn}(\nu^*_i)\left\{|\nu^*_i| - \frac{\tau\ell'(|\nu^*_i|)}{\ell'(\tau)}\right\}\mathbb{1}(|\nu^*_i| > \tau)u^\top\zeta_i\right]\right|
\end{aligned}$$

$$
\begin{aligned}
\leq & \mathbb{E}\left[\left\{|\nu_i^*| - \frac{\tau\ell'(|\nu_i^*|)}{\ell'(\tau)}\right\}\mathbb{1}(|\nu_i^*| > \tau)|u^\top\zeta_i|\right] \\
= & \mathbb{E}\left\{|\nu_i^*|\mathbb{1}(|\nu_i^*| > \tau)|u^\top\zeta_i|\right\} - \mathbb{E}\left\{\frac{\tau\ell'(|\nu_i^*|)}{\ell'(\tau)}\mathbb{1}(|\nu_i^*| > \tau)|u^\top\zeta_i|\right\} \\
\leq & \mathbb{E}\left\{|\nu_i^*|\mathbb{1}(|\nu_i^*| > \tau)|u^\top\zeta_i|\right\} \\
\leq & \tau^{-1}\mathbb{E}\left\{\nu_i^{*2}|u^\top\zeta_i|\right\} \\
= & \tau^{-1}\mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)|u^\top\zeta_i|\right\},
\end{aligned}
$$

where the second step follows from $\mathbb{1}(|\nu_i^*| \leq \tau) + \mathbb{1}(|\nu_i^*| > \tau) = 1$; the third step uses $\mathbb{E}(\nu_i^*\zeta_i) = 0$ together with the decomposition $\nu_i^* = \mathrm{sgn}(\nu_i^*)|\nu_i^*|$; the fourth step holds since $\left||\nu_i^*| - \frac{\tau\ell'(|\nu_i^*|)}{\ell'(\tau)}\right| = |\nu_i^*| - \frac{\tau\ell'(|\nu_i^*|)}{\ell'(\tau)}$ if $|\nu_i^*| \geq \tau$ by (A.11); the penultimate step uses $\mathbb{1}(|\nu_i^*| > \tau) \leq \frac{|\nu_i^*|}{\tau}$; and the final step follows by integrating out ε_i' and ε_i'', with $\bar{\varsigma}^2(\cdot)$ defined in Condition A.2 (ii).

Furthermore, using the Cauchy–Schwarz inequality, we obtain that

$$
\begin{aligned}
\left|\mathbb{E}\left\{\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)u^\top\zeta_i\right\}\right| \leq & \tau^{-1}\mathbb{E}\left\{\bar{\varsigma}(\mathrm{X}_i)|\bar{\varsigma}(\mathrm{X}_i)u^\top\zeta_i|\right\} \\
\leq & \tau^{-1}\sqrt{\mathbb{E}\{\bar{\varsigma}^2(\mathrm{X}_i)\}\mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)(u^\top\zeta_i)^2\right\}} \\
= & \tau^{-1}\sqrt{\mathbb{E}\{\bar{\varsigma}^2(\mathrm{X}_i)\}\cdot u^\top\Sigma^{-1/2}\mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)\mathrm{Z}_i\mathrm{Z}_i^\top\right\}\Sigma^{-1/2}u} \\
\leq & \frac{\sigma}{\tau}\sqrt{\sup_{u\in\mathbb{S}^{d-1}}\left[u^\top\Sigma^{-1/2}\mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)\mathrm{Z}_i\mathrm{Z}_i^\top\right\}\Sigma^{-1/2}u\right]} \\
= & \frac{\sigma}{\tau}\sqrt{\lambda_{\max}(\Sigma^{-1/2}\mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)\mathrm{Z}_i\mathrm{Z}_i^\top\right\}\Sigma^{-1/2})} \\
\triangleq & \frac{\sigma}{\tau}\varpi^{1/2},
\end{aligned} \tag{A.19}
$$

where the third step holds by the definition of ζ_i given at the beginning of Step 1.1, the fourth step uses the definition of σ in Condition A.2 (ii), the fifth step is due to Lemma A.5 (ii), and the last step arises by letting $\varpi \triangleq \lambda_{\max}(\Sigma^{-1/2}\mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)\mathrm{Z}_i\mathrm{Z}_i^\top\right\}\Sigma^{-1/2})$.

Combining (A.18) and (A.19) gives that the first term in (A.17) is upper bounded as:

$$
\text{(I)} \leq \frac{\sigma}{\tau}\varpi^{1/2} + \lambda\|\Sigma^{-1/2}\theta_{\text{reg}}^*\|_2. \tag{A.20}
$$

Step 1.2: Examining (II).

Next, we provide an upper bound on the second term (II) in (A.17). Define $\zeta^\dagger \triangleq -\sqrt{N}\Sigma^{-1/2}\left[\nabla\mathcal{L}(\theta_{\text{reg}}^*) - \mathbb{E}\left\{\nabla\mathcal{L}(\theta_{\text{reg}}^*)\right\}\right]$. Then $\text{(II)} = \left\|\frac{1}{\sqrt{N}}\zeta^\dagger\right\|_2$ by (A.17). By (A.9),

$$
\begin{aligned}
\zeta^\dagger = & \frac{1}{\sqrt{N}}\sum_{i=1}^N \mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)\Sigma^{-1/2}\mathrm{Z}_i - \lambda\sqrt{N}\Sigma^{-1/2}\theta_{\text{reg}}^* \\
& + \mathbb{E}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)\Sigma^{-1/2}\mathrm{Z}_i - \lambda\sqrt{N}\Sigma^{-1/2}\theta_{\text{reg}}^*\right\} \\
= & \frac{1}{\sqrt{N}}\sum_{i=1}^N\left[\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)\Sigma^{-1/2}\mathrm{Z}_i - \mathbb{E}\left\{\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)\Sigma^{-1/2}\mathrm{Z}_i\right\}\right] \\
\triangleq & \frac{1}{\sqrt{N}}\sum_{i=1}^N \bar{\zeta}_i,
\end{aligned} \tag{A.21}
$$

where $\bar{\zeta}_i \triangleq \mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)\Sigma^{-1/2}\mathrm{Z}_i - \mathbb{E}\left\{\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)\Sigma^{-1/2}\mathrm{Z}_i\right\} = \mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)\zeta_i - \mathbb{E}\{\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)\zeta_i\}$, with $\zeta_i \triangleq \Sigma^{-1/2}\mathrm{Z}_i$, as defined at the beginning of Step 1.1.

To establish an upper bound for $\|\zeta^\dagger\|_2$, we apply Lemma A.6 by verifying (A.2). Specifically, for any $u \in \mathbb{S}^{d-1}$ and $|t| \le \mathsf{g}$, where $\mathsf{g} \ge \sqrt{2d}$ is a constant to be specified later,

$$\begin{aligned}\mathbb{E}\{\exp(tu^\top\zeta^\dagger)\} =& \prod_{i=1}^N \mathbb{E}\left\{\exp\left(\frac{t}{\sqrt{N}}u^\top\bar{\zeta}_i\right)\right\}\\ \le& \prod_{i=1}^N \mathbb{E}\left\{1+\frac{t}{\sqrt{N}}u^\top\bar{\zeta}_i + \frac{1}{2}\left(\frac{t}{\sqrt{N}}u^\top\bar{\zeta}_i\right)^2\exp\left(\left|\frac{t}{\sqrt{N}}u^\top\bar{\zeta}_i\right|\right)\right\}\\ =& \prod_{i=1}^N \left[1+\frac{t^2}{2N}\mathbb{E}\left\{\left(u^\top\bar{\zeta}_i\right)^2\exp\left(\frac{|t|}{\sqrt{N}}|u^\top\bar{\zeta}_i|\right)\right\}\right]\end{aligned}$$

where the first step is due to the independence of ξ_i, the second step is due to the inequality $e^x \le 1 + x + \frac{x^2}{2}e^{|x|}$ for any x, and the third step holds since $\mathbb{E}(\bar{\zeta}_i) = 0$ by the definition of $\bar{\zeta}_i$ given after (A.21).

By the definition of $\bar{\zeta}_i$, (A.11), and (A.19), we obtain that for $u \in \mathbb{S}^{d-1}$,

$$\begin{aligned}|u^\top\bar{\zeta}_i| =& \left|\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)u^\top\zeta_i - \mathbb{E}\left\{\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)u^\top\zeta_i\right\}\right|\\ \le& |\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)u^\top\zeta_i| + \left|\mathbb{E}\left\{\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)u^\top\zeta_i\right\}\right|\\ \le& \tau|u^\top\zeta_i|\mathbb{1}(|\nu_i^*|\le\tau) + \frac{\tau|\nu_i|^\kappa}{\ell'(\tau)}|u^\top\zeta_i|\mathbb{1}(|\nu_i^*|>\tau) + \frac{\sigma}{\tau}\varpi^{1/2},\end{aligned}$$

and

$$\begin{aligned}(u^\top\bar{\zeta}_i)^2 =& \left[\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)u^\top\zeta_i - \mathbb{E}\left\{\mathrm{sgn}(\nu_i^*)\ell_\tau'(|\nu_i^*|)u^\top\zeta_i\right\}\right]^2\\ \le& 2\left[\{\ell_\tau'(|\nu_i^*|)u^\top\zeta_i\}^2 + \{\mathbb{E}(\ell_\tau'(|\nu_i^*|)u^\top\zeta_i)\}^2\right]\\ \le& 2\left\{\nu_i^{*2}(u^\top\zeta_i)^2 + \frac{\sigma^2}{\tau^2}\varpi\right\}.\end{aligned}$$

Consequently, for $|t| \le \mathsf{g}$ with $\mathsf{g} \ge \sqrt{2d}$ and $u \in \mathbb{S}^{d-1}$, we can further upper bound $\mathbb{E}\{\exp(tu^\top\zeta^\dagger)\}$ as:

$$\begin{aligned}&\mathbb{E}\{\exp(tu^\top\zeta^\dagger)\}\\ \le& \prod_{i=1}^N\left(1+\frac{t^2}{N}\exp\left(\frac{\sigma|t|}{\tau\sqrt{N}}\varpi^{1/2}\right)\left[\frac{\sigma^2}{\tau^2}\varpi\,\mathbb{E}\left\{\exp\left(\frac{\tau|t|}{\sqrt{N}}|u^\top\zeta_i|\right)\mathbb{1}(|\nu_i^*|\le\tau)\right\}\right.\right.\\ &\left.\left.+\frac{\sigma^2}{\tau^2}\varpi\,\mathbb{E}\left\{\exp\left(\frac{\tau|t|}{\ell'(\tau)\sqrt{N}}|u^\top\zeta_i||\nu_i^*|^\kappa\right)\mathbb{1}(|\nu_i^*|>\tau)\right\} + \mathbb{E}\left\{\nu_i^{*2}\left(u^\top\zeta_i\right)^2\exp\left(\frac{\tau|t|}{\sqrt{N}}|u^\top\zeta_i|\right)\right\}\right]\right)\\ =& \prod_{i=1}^N\left(1+\frac{t^2}{N}\exp\left(\frac{\sigma|t|}{\tau\sqrt{N}}\varpi^{1/2}\right)\left[\frac{\sigma^2}{\tau^2}\varpi\,\mathbb{E}\left\{\exp\left(\frac{\tau|t|}{\sqrt{N}}|u^\top\zeta_i|\right)\right\}\right.\right.\\ &\left.\left.+\frac{\sigma^2}{\tau^2}\varpi\,\mathbb{E}\left\{\exp\left(\frac{\tau|t|}{\ell'(\tau)\sqrt{N}}|u^\top\zeta_i||\nu_i^*|^\kappa\right)\mathbb{1}(|\nu_i^*|>\tau)\right\} + \mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)\left(u^\top\zeta_i\right)^2\exp\left(\frac{\tau|t|}{\sqrt{N}}|u^\top\zeta_i|\right)\right\}\right]\right)\\ \le& \prod_{i=1}^N\left[1+\frac{t^2}{N}\exp\left\{\frac{\sqrt{2}(d+s)}{N}\varpi^{1/2}\right\}\left\{\frac{(d+s)\varpi}{N}\underbrace{\left[\mathbb{E}\left\{\exp\left(\sqrt{2}\sigma|u^\top\zeta_i|\right)\right\} + \mathbb{E}\left\{\exp\left(\sqrt{2}\sigma|u^\top\zeta_i||\nu_i^*|^\kappa\right)\right\}\right]}_{\text{(II)-a}}\right.\right.\\ &\left.\left.+\underbrace{\mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)\left(u^\top\zeta_i\right)^2\exp\left(\sqrt{2}\sigma|u^\top\zeta_i|\right)\right\}}_{\text{(II)-b}}\right\}\right],\end{aligned} \tag{A.22}$$

where the second step holds by integrating out the noise variables ε_i' and ε_i'', with $\bar{\varsigma}^2(\cdot)$ given in Section A.2 (ii), and the last step is true when $\frac{\tau|t|}{\sqrt{N}} \le \sqrt{2}\sigma$ and $\frac{\tau|t|}{\ell'(\tau)\sqrt{N}} \le \sqrt{2}\sigma$, which is ensured by letting $|t| \le \mathsf{g} = \sqrt{2(d+s)}$ and $\max(\tau, \tau/\ell'(\tau)) \le \sigma\sqrt{N/(d+s)}$ as required in Theorem A.15.

We next identify upper bounds of (II)-a and (II)-b in (A.22). For (II)-a with any $u \in \mathbb{S}^{d-1}$, we have

$$\begin{aligned}\text{(II)-a} &\leq \sup_{v\in\mathbb{S}^{d-1}} \left[\mathbb{E}\left\{\exp\left(\sqrt{2}\sigma|u^\top\zeta_i|\right)\right\} + \mathbb{E}\left\{\exp\left(\sqrt{2}\sigma|u^\top\zeta_i||\nu_i^*|^\kappa\right)\right\}\right] \\ &\leq \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\left[\exp\left\{\sqrt{2}\sigma(v^\top\zeta_i)\right\}\right] + \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\left[\exp\left\{\sqrt{2}\sigma v^\top\left(\zeta_i|\nu_i^*|^\kappa\right)\right\}\right] \leq \mathcal{K}_1 \end{aligned} \tag{A.23}$$

for some constant $\mathcal{K}_1 > 0$, where the second step holds by the symmetry of $\mathbb{S}^{d-1}$ and the subadditivity of supremum, and the last step is due to the sub-Gaussian assumption in Condition A.2 (iii) and the equivalence properties in Lemma A.7, with $\mathcal{K}_1$ denoting a constant depending on σ in Condition A.2 (ii), α_2 in Condition A.2 (iii) (a), and α_3 in Condition A.2 (iii) (b).

For (II)-b with any $u \in \mathbb{S}^{d-1}$, we have that

$$\begin{aligned}\text{(II)-b} \leq& \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\left\{\bar{\varsigma}^2(\mathrm{X}_i)\left(v^\top\zeta_i\right)^2\exp\left(\sqrt{2}\sigma|v^\top\zeta_i|\right)\right\} \\ \leq& \sup_{v\in\mathbb{S}^{d-1}} \Bigg\{ \left[\mathbb{E}\{\bar{\varsigma}^{2+\iota_1}(\mathrm{X}_i)\}\right]^{\frac{2}{2+\iota_1}} \left[\mathbb{E}\{(v^\top\zeta_i)^{2(4/\iota_1+2)}\}\right]^{\iota_1/(4+2\iota_1)} \\ &\qquad \left(\mathbb{E}[\exp\{\sqrt{2}\sigma(4/\iota_1+2)|v^\top\zeta_i|\}]\right)^{\iota_1/(4+2\iota_1)} \Bigg\} \\ \leq& \left[\mathbb{E}\{\bar{\varsigma}^{2+\iota_1}(\mathrm{X}_i)\}\right]^{\frac{2}{2+\iota_1}} \cdot \sup_{v\in\mathbb{S}^{d-1}} \left[\mathbb{E}\{(v^\top\zeta_i)^{2(4/\iota_1+2)}\}\right]^{\iota_1/(4+2\iota_1)} \\ &\cdot \sup_{v\in\mathbb{S}^{d-1}} \left(\mathbb{E}[\exp\{\sqrt{2}\sigma(4/\iota_1+2)(v^\top\zeta_i)\}]\right)^{\iota_1/(4+2\iota_1)} \\ \leq& \mathcal{K}_2, \end{aligned} \tag{A.24}$$

for some constant $\mathcal{K}_2 > 0$, where the second step is due to holder's inequality, with ι_1 given in Condition A.2 (ii), and the last step holds by Condition A.2 (ii) and the sub-Gaussian assumption in Condition A.2 (iii), as well as the equivalence properties in Lemma A.7, with the constant $\mathcal{K}_2$ depending on α_1, ι_1, and σ in Condition A.2 (ii) and α_2 in Condition A.2 (iii).

Applying (A.23) and (A.24) to (A.22), we obtain that, for $|t| \leq \mathsf{g} = \sqrt{2(d+s)}$ and $u \in \mathbb{S}^{d-1}$,

$$\mathbb{E}\{\exp(tu^\top\zeta^\dagger)\} \leq \prod_{i=1}^N \left(1 + \frac{t^2}{N}\exp\left(\frac{\sqrt{2}(d+s)}{N}\varpi^{1/2}\right)\left\{\frac{(d+s)\varpi}{N}\,\mathcal{K}_1 + \mathcal{K}_2\right\}\right) \leq \exp(\mathcal{K}^\dagger t^2/2),$$

where $\mathcal{K}^\dagger \triangleq \max\left\{2e^{\sqrt{2}\varpi^{1/2}}(\varpi\mathcal{K}_1 + \mathcal{K}_2), 1.1\right\}$, and the last step is due to the inequality $1+x \leq e^x$, along with the assumption $N > d+s$ in Theorem A.15. Hence, $\mathcal{K}^\dagger$ is a constant depending on α_1, ι_1, and σ in Condition A.2 (ii), α_2 in Condition A.2 (iii) (a), α_3 in Condition A.2 (iii) (b), and ϖ defined after (A.19). By rescaling $\zeta^\dagger$ and g, we further obtain that, for $|t| \leq \bar{\mathsf{g}} \triangleq \sqrt{\mathcal{K}^\dagger}\mathsf{g} = \sqrt{2\mathcal{K}^\dagger(d+s)}$, and $u \in \mathcal{S}^{d-1}$,

$$\mathbb{E}\left[\exp\left\{tu^\top\left(\zeta^\dagger/\sqrt{\mathcal{K}^\dagger}\right)\right\}\right] \leq e^{t^2/2}.$$

Consequently, by Lemma A.6 with $V_0 = D_0 = I_{d\times d}$, we obtain that for $\sqrt{2d}/18 < \mathtt{x}^* \leq \mathtt{c}_4$,

$$\mathbb{P}\left\{\left\|\zeta^\dagger\right\|_2^2 > \mathcal{K}^\dagger(d + 6\mathtt{x}^*)\right\} = \mathbb{P}\left(\left\|\zeta^\dagger/\sqrt{\mathcal{K}^\dagger}\right\|_2^2 > d + 6\mathtt{x}^*\right) \leq 2e^{-\mathtt{x}^*} + 8.4e^{-\mathtt{c}_4}, \tag{A.25}$$

where $\mathtt{c}_4 = \frac{1}{2}\left(\bar{\mathsf{g}}^2 - d\log\frac{d+\bar{\mathsf{g}}^2}{d}\right)$. By taking the derivative of $\mathtt{c}_4$ with respect to $\bar{\mathsf{g}}$, we obtain that $\frac{d\mathtt{c}_4}{d\bar{\mathsf{g}}} = \frac{\bar{\mathsf{g}}^3}{d+\bar{\mathsf{g}}^2} > 0$, and hence, $\mathtt{c}_4$ is increasing in $\bar{\mathsf{g}}$. Therefore, since $\bar{\mathsf{g}} = \sqrt{2\mathcal{K}^\dagger(d+s)} \geq \sqrt{2.2(d+s)}$, we obtain that

$$\mathtt{c}_4 \geq \frac{1}{2}\left\{2.2(d+s) - d\log\left(3.2 + \frac{2.2s}{d}\right)\right\} \geq \frac{1}{2}\left\{2.2(d+s) - d\left(0.2\cdot\frac{s}{d} + 1.8\right)\right\} = s + 0.2d,$$

where the second inequality holds since $\log(3.2+2.2x) \le 0.2x+1.8$ for $x \ge 0$. By taking $\mathrm{x}^* = s + 0.2d$ in (A.25), we obtain that, when $d > 10$, with probability at least $1-3e^{-s}$, $\left\|\zeta^\dagger\right\|_2^2 \le \mathcal{K}^\dagger(2.2d+6s)$, and consequently,

$$\text{(II)} = \frac{1}{\sqrt{N}}\left\|\zeta^\dagger\right\|_2 \le \sqrt{\frac{\mathcal{K}^\dagger(2.2d+6s)}{N}}. \tag{A.26}$$

Combining (A.17), (A.20), and (A.26), we obtain that, with probability at least $1-3e^{-s}$

$$\left\|\Sigma^{-1/2}\nabla\mathcal{L}(\theta^*_{\text{reg}})\right\|_2 \le \frac{\sigma}{\tau}\varpi^{1/2} + \lambda\|\Sigma^{-1/2}\theta^*_{\text{reg}}\|_2 + \sqrt{\frac{\mathcal{K}^\dagger(2.2d+6s)}{N}}. \tag{A.27}$$

Step 2: Lower bounding $\lambda_{\min}\left(\Sigma^{-1/2}\nabla^2\mathcal{L}(\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta))\Sigma^{-1/2}\right)$ in (A.16).

Let m^* be a given constant, and we can choose $\eta \in [0,1]$ such that

$$\left\|\Sigma^{1/2}\left\{\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}}\right\}\right\|_2 \le m^* \tag{A.28}$$

Indeed, if $\left\|\Sigma^{1/2}\left(\widehat{\theta}^{\tau,\lambda}_{\text{reg}} - \theta^*_{\text{reg}}\right)\right\|_2 \le m^*$, then we take $\eta = 1$; otherwise, we can always choose some $\eta < 1$ such that $\left\|\Sigma^{1/2}\left\{\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}}\right\}\right\|_2 = m^*$. Since $\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)$ is a convex combination of $\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)$ and θ^*_{reg}, then replacing $\widehat{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)$ in (A.28) with $\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta)$ yields $\left\|\Sigma^{1/2}\left\{\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta) - \theta^*_{\text{reg}}\right\}\right\|_2 \le m^*$. Consequently,

$$\begin{aligned}
&\lambda_{\min}\left(\Sigma^{-1/2}\nabla^2\mathcal{L}(\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta))\Sigma^{-1/2}\right)\\
\ge&\inf_{\theta_{\text{reg}}:\left\|\Sigma^{1/2}\left(\theta_{\text{reg}}-\theta^*_{\text{reg}}\right)\right\|_2\le m^*}\lambda_{\min}\left(\Sigma^{-1/2}\nabla^2\mathcal{L}(\theta_{\text{reg}})\Sigma^{-1/2}\right)\\
=&\inf_{\theta_{\text{reg}}:\left\|\Sigma^{1/2}\left(\theta_{\text{reg}}-\theta^*_{\text{reg}}\right)\right\|_2\le m^*}\lambda_{\min}\left(\Sigma^{-1/2}\left\{\frac{1}{N}\sum_{i=1}^N \mathrm{Z}_i\mathrm{Z}_i^\top\ell''_\tau(|\nu_i|)+\lambda I_{d\times d}\right\}\Sigma^{-1/2}\right)\\
=&\inf_{\theta_{\text{reg}}:\left\|\Sigma^{1/2}\left(\theta_{\text{reg}}-\theta^*_{\text{reg}}\right)\right\|_2\le m^*}\lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^N\zeta_i\zeta_i^\top\mathbb{1}(|\nu_i|\le\tau)+\frac{1}{N}\sum_{i=1}^N\zeta_i\zeta_i^\top\frac{\tau\ell''(|\nu_i|)}{\ell'(\tau)}\mathbb{1}(|\nu_i|>\tau)+\lambda\Sigma^{-1}\right)\\
\ge&\inf_{\theta_{\text{reg}}:\left\|\Sigma^{1/2}\left(\theta_{\text{reg}}-\theta^*_{\text{reg}}\right)\right\|_2\le m^*}\lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^N\zeta_i\zeta_i^\top\mathbb{1}(|\nu_i|\le\tau)\right)+\lambda_{\min}(\lambda\Sigma^{-1})\\
=&\inf_{\theta_{\text{reg}}:\left\|\Sigma^{1/2}\left(\theta_{\text{reg}}-\theta^*_{\text{reg}}\right)\right\|_2\le m^*}\inf_{u\in\mathbb{S}^{d-1}}\left\{\frac{1}{N}\sum_{i=1}^N(u^\top\zeta_i)^2\mathbb{1}(|\nu_i|\le\tau)\right\}+\lambda_{\min}(\lambda\Sigma^{-1}),
\end{aligned}\tag{A.29}$$

where the second step is due to (A.10), the third step comes from (A.12) and the definition $\zeta_i = \Sigma^{-1/2}\mathrm{Z}_i$ at the beginning of Step 1.1, the fourth step is due to Lemma A.5 (iii) and the assumpion that $\ell(\cdot)$ is convex in Theorem A.15, and the last step holds by Lemma A.5 (ii). For any θ_{reg} satisfying $\left\|\Sigma^{1/2}\left(\theta_{\text{reg}}-\theta^*_{\text{reg}}\right)\right\|_2 \le m^*$, we have that

$$\begin{aligned}
|\nu_i| =&|\widetilde{\mathrm{Y}}_i - \langle\mathrm{Z}_i,\theta_{\text{reg}}\rangle| = |\langle\mathrm{Z}_i,\theta^*_{\text{reg}}\rangle + \nu_i^* - \langle\mathrm{Z}_i,\theta_{\text{reg}}\rangle|\\
\le&|\mathrm{Z}_i^\top(\theta^*_{\text{reg}}-\theta_{\text{reg}})| + |\nu_i^*| = |\zeta_i^\top\Sigma^{-1/2}(\theta^*_{\text{reg}}-\theta_{\text{reg}})| + |\nu_i^*|\\
\le&\|\zeta_i\|_2\|\Sigma^{-1/2}(\theta^*_{\text{reg}}-\theta_{\text{reg}})\|_2 + |\nu_i^*| \le m^*\|\zeta_i\|_2 + |\nu_i^*|,
\end{aligned}\tag{A.30}$$

where ν_i is defined after (A.8), and $\nu_i^* = \varepsilon' + \varsigma(\mathrm{X},\mathrm{Y})\varepsilon''$, as defined at the beginning of Step 1.1. Therefore, by (A.29) and (A.30), we further obtain that

$$\begin{aligned}
&\lambda_{\min}\left(\Sigma^{-1/2}\nabla^2\mathcal{L}(\widetilde{\theta}^{\tau,\lambda}_{\text{reg}}(\eta))\Sigma^{-1/2}\right)\\
\ge&\inf_{u\in\mathbb{S}^{d-1}}\left\{\frac{1}{N}\sum_{i=1}^N(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\le\tau)\right\}+\lambda_{\min}(\lambda\Sigma^{-1})
\end{aligned}$$

$$
\begin{aligned}
&\geq \inf_{u\in\mathbb{S}^{d-1}}\left[\frac{1}{N}\sum_{i=1}^{N}(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)\}\right]\\
&\quad+\inf_{u\in\mathbb{S}^{d-1}}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)\}\right]+\lambda_{\min}(\lambda\Sigma^{-1})\\
&=\lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^{N}\zeta_i\zeta_i^\top\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{\zeta_i\zeta_i^\top\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)\}\right)\\
&\quad+\inf_{u\in\mathbb{S}^{d-1}}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)\}\right]+\lambda_{\min}(\lambda\Sigma^{-1})\\
&\geq \inf_{u\in\mathbb{S}^{d-1}}\underbrace{\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)\}\right]}_{\text{(III)-a}}+\lambda_{\min}(\lambda\Sigma^{-1})\\
&\quad-\underbrace{\left\|\frac{1}{N}\sum_{i=1}^{N}\zeta_i\zeta_i^\top\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{\zeta_i\zeta_i^\top\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)\}\right\|_2}_{\text{(III)-b}}, \quad \text{(A.31)}
\end{aligned}
$$

where the third step is due to Lemma A.5 (ii), and last step is due to Lemma A.5 (i).

Next, we examine (III)-a. Specifically, for any $u\in\mathbb{S}^{d-1}$,

$$
\begin{aligned}
\text{(III)-a}=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|\leq\tau)\}\\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\}-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|>\tau)\}\\
=&\frac{1}{N}\sum_{i=1}^{N}u^\top\Sigma^{-1/2}\mathbb{E}\{Z_iZ_i^\top\}\Sigma^{-1/2}u-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|>\tau)\}\\
=&1-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2\mathbb{1}(m^*\|\zeta_i\|_2+|\nu_i^*|>\tau)\}\\
\geq&1-\frac{1}{\tau N}\sum_{i=1}^{N}\mathbb{E}\{(u^\top\zeta_i)^2(m^*\|\zeta_i\|_2+|\nu_i^*|)\}\\
=&1-\frac{1}{\tau N}\sum_{i=1}^{N}\left[m^*\mathbb{E}\{(u^\top\zeta_i)^2\|\zeta_i\|_2\}+\mathbb{E}\{(u^\top\zeta_i)^2|\nu_i^*|\}\right]\\
\geq&1-\frac{1}{\tau N}\sum_{i=1}^{N}\left[m^*\sqrt{\mathbb{E}\{(u^\top\zeta_i)^4\}}\sqrt{\mathbb{E}(\zeta_i^\top\zeta_i)}+\sqrt{\mathbb{E}\{(u^\top\zeta_i)^4\}}\sqrt{\mathbb{E}(\nu_i^{*2})}\right]\\
\geq&1-\frac{1}{\tau N}\sum_{i=1}^{N}4\alpha_1^2\left[m^*\sqrt{\mathbb{E}(\zeta_i^\top\zeta_i)}+\sigma\right],
\end{aligned}
$$

where the fourth step is due to the definition of Σ given after (A.14) and the fact that $u\in\mathbb{S}^{d-1}$, the fifth step is due to the fact that $\mathbb{1}(a>b)\leq\frac{a}{b}$ for any $a>0$ and $b>0$, the penultimate step is due to the Cauchy–Schwarz inequality, and the last step comes from Condition A.2 (iii) and the Lemma A.7 (ii). Since

$$
\begin{aligned}
\mathbb{E}\{\zeta_i^\top\zeta_i\}=&\mathbb{E}(Z_i^\top\Sigma^{-1}Z_i)=\mathbb{E}\{\text{tr}(Z_i^\top\Sigma^{-1}Z_i)\}=\mathbb{E}\{\text{tr}(\Sigma^{-1}Z_iZ_i^\top)\}\\
=&\text{tr}\{\mathbb{E}(\Sigma^{-1}Z_iZ_i^\top)\}=\text{tr}(I_{d\times d})=d,
\end{aligned}
$$

we further obatain that

$$\text{(III)-a} \geq 1 - 4\alpha_1^2 \tau^{-1}(m^*\sqrt{d} + \sigma). \tag{A.32}$$

Now we establish a lower bound for (III)-b in (A.31) using Lemma A.9. Specifically, write

$$\frac{1}{N}\sum_{i=1}^{N} \zeta_i \zeta_i^\top \mathbb{1}(m^*\|\zeta_i\|_2 + |\nu_i^*| \leq \tau) \triangleq \frac{1}{N} A^\top A,$$

where $A = (A_1, \ldots, A_N)^\top$ with $A_i = \zeta_i \mathbb{1}(m^*\|\zeta_i\|_2 + |\nu_i^*| \leq \tau)$ for $i \in [N]$. By Condition A.2 (iii) and Lemma A.9, we obtain that with probability at least $1 - 2e^{-s}$,

$$\text{(III)-b} \leq \max(\epsilon, \epsilon^2), \text{ with } \epsilon = \mathcal{K}_4 \sqrt{\frac{d+s}{N}} \tag{A.33}$$

for some constant $\mathcal{K}_4 > 0$ that depends only on α_2 in Condition A.2 (iii).

Combining (A.31), (A.32), and (A.33), we obtain that with probability at least $1 - 2e^{-s}$,

$$\lambda_{\min}\left(\Sigma^{-1/2}\nabla^2 \mathcal{L}(\widetilde{\theta}_{\text{reg}}^{\tau,\lambda}(\eta))\Sigma^{-1/2}\right) \geq 1 - 4\alpha_1^2\tau^{-1}(m^*\sqrt{d} + \sigma) - \max(\epsilon, \epsilon^2) \geq \frac{1}{4}, \tag{A.34}$$

where the last inequality holds if $\tau \geq 8\alpha_1^2(m^*\sqrt{d} + \sigma)$ and $N \geq 16\mathcal{K}_4^2(d+s)$, which is guaranteed by taking $\tau \geq \mathscr{C}_1(\sqrt{d} + \sigma)$ and $N \geq \mathscr{C}_2(d+s)$ as required by Theorem A.15.

Combining (A.16), (A.27), and (A.34), we obtain that with probability at least $1 - 5e^{-s}$,

$$\left\|\Sigma^{1/2}\left\{\widehat{\theta}_{\text{reg}}^{\tau,\lambda}(\eta) - \theta_{\text{reg}}^*\right\}\right\|_2 \leq 4\left\{\frac{\sigma}{\tau}\varpi^{1/2} + \lambda\|\Sigma^{-1/2}\theta_{\text{reg}}^*\|_2 + \sqrt{\frac{\mathcal{K}^\dagger(2.2d + 6s)}{N}}\right\} < m^*,$$

if we take, for example, $m^* = 4.1\left\{\frac{\sigma}{\tau}\varpi^{1/2} + \lambda\|\Sigma^{-1/2}\theta_{\text{reg}}^*\|_2 + \sqrt{\frac{\mathcal{K}^\dagger(2.2d+6s)}{N}}\right\}$. Consequently, $\widehat{\theta}_{\text{reg}}^{\tau,\lambda}(\eta)$ is an interior point of the ball with center θ_{reg}^* and radius m^*, and thus, by the construction of $\widehat{\theta}_{\text{reg}}^{\tau,\lambda}(\eta)$ at the beginning of Step 2, we can take $\eta = 1$. Therefore, under the assumptions of Theorem A.15, with probability at least $1 - 5e^{-s}$,

$$\left\|\widehat{\theta}_{\text{reg}}^{\tau,\lambda} - \theta_{\text{reg}}^*\right\|_\Sigma \triangleq \left\|\Sigma^{1/2}\left\{\widehat{\theta}_{\text{reg}}^{\tau,\lambda}(\eta) - \theta_{\text{reg}}^*\right\}\right\|_2 \leq \mathscr{C}^\dagger\sqrt{\frac{d+s}{N}}$$

for some constant $\mathscr{C}^\dagger > 0$. Hence, the proof of Theorem A.15 (i) is completed.

A.5 Proof of Theorem 4.2 (ii) (Theorem A.15 (ii))

The prediction error, defined by $\text{err} \triangleq \left|\langle \text{Z}_{\text{new}}, \widehat{\theta}_{\text{reg}}^{\tau,\lambda}\rangle - \langle \text{Z}_{\text{new}}, \theta_{\text{reg}}^*\rangle\right|$, can be upper bounded as below:

$$\begin{aligned}
\text{err} &= \left|(\Sigma^{-1/2}\text{Z}_{\text{new}})^\top \Sigma^{1/2}(\widehat{\theta}_{\text{reg}}^{\tau,\lambda} - \theta_{\text{reg}}^*)\right| \\
&= \left|\langle \Sigma^{-1/2}\text{Z}_{\text{new}}, \frac{\Sigma^{1/2}(\widehat{\theta}_{\text{reg}}^{\tau,\lambda} - \theta_{\text{reg}}^*)}{\|\widehat{\theta}_{\text{reg}}^{\tau,\lambda} - \theta_{\text{reg}}^*\|_\Sigma}\rangle\right| \cdot \|\widehat{\theta}_{\text{reg}}^{\tau,\lambda} - \theta_{\text{reg}}^*\|_\Sigma \\
&\leq \sup_{u \in \mathbb{S}^{d-1}} |\langle \Sigma^{-1/2}\text{Z}_{\text{new}}, u\rangle| \cdot \|\widehat{\theta}_{\text{reg}}^{\tau,\lambda} - \theta_{\text{reg}}^*\|_\Sigma.
\end{aligned}$$

By Condition A.2 (iii) and Lemma A.7 (ii), for any $s > 0$, with probability at least $1 - 2e^{-s}$,

$$\sup_{u \in \mathbb{S}^{d-1}} |\langle \Sigma^{-1/2}\text{Z}_{\text{new}}, u\rangle| \leq \mathcal{K}_5\sqrt{s}$$

for some constant $\mathcal{K}_5 > 0$, which depends only on α_2 in Condition A.2 (iii). Consequently, combining with Theorem A.15 (i), we obtain that, with probability at least $1 - 7e^{-s}$,

$$\text{err} \leq \mathscr{C}^\ddagger\sqrt{\frac{ds + s^2}{N}}$$

for some constant $\mathscr{C}^\ddagger > 0$, which completes the proof of Theorem A.15 (ii).

A.6 Proof of Proposition 4.4

Proof. We verify that ℓ_τ satisfies the conditions in Lemma A.12. By Lemma A.13, ℓ_τ is convex. Then it suffices to verify that ℓ'_τ is 1-Lipschitz. We consider the following three cases.

(i) For $0 \le s \le t \le \tau$, this is immediate since $\ell'_\tau(t) - \ell'_\tau(s) = t - s$.

(ii) For $\tau \le s \le t$, since $\ell'(t)/t$ is non-increasing and $s \le t$, we have

$$\frac{\ell'(t)}{t} \le \frac{\ell'(s)}{s},$$

which rearranges to

$$\ell'(t) - \ell'(s) \le \frac{t-s}{s}\ell'(s).$$

Since $\ell'(t)/t$ is non-increasing and $s \ge \tau$, we also have $\ell'(s)/s \le \ell'(\tau)/\tau$, so $\ell'(s) \le s\ell'(\tau)/\tau$. Therefore,

$$\ell'(t) - \ell'(s) \le \frac{t-s}{s} \cdot \frac{s\,\ell'(\tau)}{\tau} = \frac{(t-s)\ell'(\tau)}{\tau}.$$

Multiplying both sides by $\tau/\ell'(\tau)$ gives

$$\frac{\tau}{\ell'(\tau)}\{\ell'(t) - \ell'(s)\} \le t - s,$$

i.e., $\ell'_\tau(t) - \ell'_\tau(s) \le t - s$.

(iii) For the cross-region case $0 \le s \le \tau \le t$,

$$\ell'_\tau(t) - \ell'_\tau(s) = \big\{\ell'_\tau(t) - \ell'_\tau(\tau)\big\} + \{\ell'_\tau(\tau) - \ell'_\tau(s)\} \le (t-\tau) + (\tau - s) = t - s.$$

where the inequality holds by applying the upper bounds in cases (i) and (ii).

Since ℓ_τ is even, the same Lipschitz bound extends to the whole real line by symmetry. Thus ℓ_τ is convex and has 1-Lipschitz derivative. Therefore, by Lemma A.12, we obtain

$$\ell_\tau(\nu) = \inf_{\xi \in \mathbb{R}} \left\{ \frac{1}{2}\xi^2 + \phi_\tau(\nu - \xi) \right\}, \tag{A.35}$$

where $\phi_\tau = \psi_\tau^*$, with $\psi_\tau(w) \triangleq \ell_\tau^*(w) - \frac{1}{2}w^2$. This proves the infimal convolution representation. It remains to derive the parametric form of ϕ_τ.

Fix $t \ge \tau$ and set

$$w = \ell'_\tau(t) = \frac{\tau}{\ell'(\tau)}\ell'(t), \qquad u = t - w = t - \frac{\tau\ell'(t)}{\ell'(\tau)}. \tag{A.36}$$

Here we restrict to $t \ge \tau$, since for $0 \le t \le \tau$, $\ell_\tau(t) = t^2/2$ and hence $\ell'_\tau(t) = t$, which implies $u = t - \ell'_\tau(t) = 0$. Therefore, all nontrivial values $u > 0$ arise from the regime $t \ge \tau$.

By convexity and differentiability of ℓ_τ, we have $w \in \partial\ell_\tau(t)$, so the Fenchel–Young equality gives

$$\ell_\tau^*(w) = tw - \ell_\tau(t). \tag{A.37}$$

Next, since $\phi_\tau^*(w) = \ell_\tau^*(w) - \frac{1}{2}w^2$, we obtain

$$\partial\phi_\tau^*(w) = \partial\ell_\tau^*(w) - w. \tag{A.38}$$

Since $w \in \partial\ell_\tau(t)$ by (A.36), we have $t \in \partial\ell_\tau^*(w)$ by conjugacy, and hence

$$u = t - w \in \partial\ell_\tau^*(w) - w = \partial\phi_\tau^*(w).$$

Another application of the Fenchel–Young equality yields

$$\phi_\tau(u) + \phi_\tau^*(w) = uw. \tag{A.39}$$

Therefore,

$$\phi_\tau(u) = uw - \phi_\tau^*(w) \tag{A.40}$$

$$= uw - \ell_\tau^*(w) + \frac{1}{2}w^2 \tag{A.41}$$

$$= uw - \{tw - \ell_\tau(t)\} + \frac{1}{2}w^2 \tag{A.42}$$

$$= \ell_\tau(t) - \frac{1}{2}w^2, \tag{A.43}$$

where the first step follows from (A.39), the second step comes from the definition of ϕ_τ^*, and the third step is due to (A.37).

Substituting $\ell_\tau(t) = \frac{\tau}{\ell'(\tau)}\{\ell(t) - \ell(\tau)\} + \frac{\tau^2}{2}$ and $w = \frac{\tau\ell'(t)}{\ell'(\tau)}$ into (A.40) for $t \geq \tau$ yields

$$\phi_\tau(u) = \frac{\tau}{\ell'(\tau)}\big\{\ell(t) - \ell(\tau)\big\} + \frac{\tau^2}{2} - \frac{1}{2}\left\{\frac{\tau\ell'(t)}{\ell'(\tau)}\right\}^2.$$

It remains to show that the mapping $t \mapsto u(t)$ is injective and covers all $u \geq 0$. Define

$$u(t) = t - \frac{\tau}{\ell'(\tau)}\ell'(t), \qquad t \geq \tau.$$

Let $r(t) \triangleq \ell'(t)/t$ and $r_\tau \triangleq r(\tau) = \ell'(\tau)/\tau > 0$. Then

$$u(t) = t\left\{1 - \frac{r(t)}{r_\tau}\right\}.$$

By assumption, $r(t)$ is strictly decreasing on $[\tau, \infty)$, and hence $u(\tau) = 0$ and $u(t) > 0$ for all $t > \tau$.

For any $\tau \leq s < t$, we have $r(t) < r(s) \leq r_\tau$. Therefore,

$$\ell'(t) - \ell'(s) = tr(t) - sr(s) < (t-s)r(s) \leq (t-s)r_\tau.$$

Consequently,

$$u(t) - u(s) = (t-s) - \frac{\ell'(t) - \ell'(s)}{r_\tau} > 0.$$

Thus $u(t)$ is strictly increasing on $[\tau, \infty)$ and hence injective.

To establish surjectivity, fix any $t_0 > \tau$. Since $r(t)$ is strictly decreasing, $r(t_0) < r_\tau$. For all $t \geq t_0$,

$$u(t) = t\left\{1 - \frac{r(t)}{r_\tau}\right\} \geq t\left\{1 - \frac{r(t_0)}{r_\tau}\right\}.$$

Since $1 - r(t_0)/r_\tau > 0$, it follows that $u(t) \to \infty$ as $t \to \infty$. Together with the continuity of $u(t)$ and $u(\tau) = 0$, this shows that $u(t)$ maps $[\tau, \infty)$ onto $[0, \infty)$.

Therefore, $t \mapsto u(t)$ is a bijection from $[\tau, \infty)$ onto $[0, \infty)$, ensuring that the parametric characterization of $\phi_\tau(u)$ is well-defined for every $u \geq 0$.

Finally, since ℓ_τ is even, for any w we have

$$\ell_\tau^*(-w) = \sup_x\{-wx - \ell_\tau(x)\} = \sup_x\{w(-x) - \ell_\tau(-x)\} = \sup_x\{wx - \ell_\tau(x)\} = \ell_\tau^*(w).$$

Thus, ℓ_τ^* is even. Hence $\phi_\tau^*(w) = \ell_\tau^*(w) - w^2/2$ is even, and therefore $\phi_\tau = (\phi_\tau^*)^*$ is also even, giving $\phi_\tau(-u) = \phi_\tau(u)$. At $t = \tau$, we have $u = 0$ and thus $\phi_\tau(0) = 0$. This completes the proof. □

B Experiment Details

B.1 Experiment Setup

Metric Computation. Let y_i and $\hat{y}_i$ denote the ground-truth and predicted labels for the i-th sample, and let N be the total number of test samples. We employ the following three commonly

used evaluation metrics to assess the performance for regression tasks: MAE $= \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$; RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$; and $R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$. Here, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$ denote the sample means of the ground-truth and predicted labels, respectively. MAE and RMSE measure prediction accuracy from different perspectives, and R^2 characterizes the goodness-of-fit between predictions and ground-truth values.

Dataset information. We evaluate our method on four widely used real-world regression benchmarks spanning facial analysis, human pose estimation, and medical imaging.

- **UTKFace** [19] is a large-scale facial dataset containing over $N = 20,000$ images annotated with age, gender, and ethnicity. In our experiments, face images are resized to $p = 3 \times 224 \times 224$ and used as inputs X, age annotations serve as the regression targets Y, and synthetic Gaussian noise is added to generate noisy labels $\widetilde{\mathrm{Y}}$. A ResNet-50 backbone is used as the feature extractor, producing $d = 2048$-dimensional feature representations.
- **Biwi Kinect** [21] is a head pose estimation dataset consisting of RGB-D images captured with a Microsoft Kinect sensor, annotated with continuous head pose angles including yaw, pitch, and roll. The dataset contains approximately $N = 15,000$ images from 20 subjects (6 females and 14 males), where four subjects were recorded twice to increase pose variability and robustness. In our experiments, RGB images are resized to $p = 3 \times 224 \times 224$ and used as inputs X, the head pose angles serve as the regression targets Y, and synthetic Gaussian noise is added to generate noisy labels $\widetilde{\mathrm{Y}}$. A ResNet-50 backbone is used as the feature extractor, producing $d = 2048$-dimensional feature representations.
- **APPA-REAL** [20] contains 7,591 facial images annotated with both real age and apparent age, where apparent age refers to the age perceived by human annotators rather than the chronological age of the subject. The dataset is split into 4,113 training, 1,500 validation, and 1,978 test images. In our experiments, face images are resized to $p = 3 \times 224 \times 224$ and used as inputs X, apparent age annotations are treated as noisy labels $\widetilde{\mathrm{Y}}$, and real age labels are used as the ground-truth targets Y for evaluation. A ResNet-50 backbone is used as the feature extractor, producing $d = 2048$-dimensional feature representations.
- For **Bone-Age** assessment, we combine two datasets: **RSNA** [22] and **DHA** (Digital Hand Atlas) [23]. The RSNA dataset contains 12,611 training and 1,425 validation hand X-ray images with manually annotated bone ages, and is used for model training. The DHA dataset includes 1,400 left hand and wrist radiographs of children, accompanied by demographic information and radiologists' assessments based on visual comparison with an atlas and computer-aided diagnosis (CAD). In our experiments, hand X-ray images, resized to $p = 224 \times 224$, are used as inputs X, bone age annotations from RSNA are treated as noisy labels $\widetilde{\mathrm{Y}}$ for model training, and radiologist-assessed bone ages from DHA are used as ground-truth targets Y for evaluation. A DenseNet-121 backbone is used as the feature extractor, producing $d = 1024$-dimensional feature representations.

Noise Simulation. For UTKFace and Biwi Kinect, we generate synthetic Gaussian label noise using the following procedures:

1. Let $\mathcal{D}_{\mathrm{y}} \triangleq \{\mathrm{Y}_i\}_{i=1}^{N}$ denote the set of clean labels, and let $\mathrm{y}_{\min}$ and $\mathrm{y}_{\max}$ be the minimum and maximum label values in the dataset.
2. Uniformly partition the interval $[y_{\min}, y_{\max}]$ into M disjoint sub-intervals. In all experiments, we set $M = 4000$, and let $\mathrm{std}_{\max} \in \{30, 60, 90\}$ to control the maximum noise magnitude.
3. For each sample, uniformly draw an integer variance σ^2 from $[\mathrm{std}_{\max}]$ and then sample a noise offset from $\mathcal{N}(0, \sigma^2)$. The resulting offset determines an interval index. Let $[y_l, y_u]$ denote the corresponding sub-interval, and define $\Omega \triangleq \mathcal{D}_y \cap [y_l, y_u]$. Noisy label $\widetilde{Y}_i$ is taken as a value drawn from the uniform distribution over Ω.

This noise generation strategy exhibits several desirable properties. First, by sampling noise magnitudes at the instance level, it introduces heteroscedastic and sample-varying perturbations, which

more closely resemble real-world annotation noise than i.i.d. Gaussian corruption. Second, sampling noisy labels from nearby intervals preserves the ordered and metrized structure of the continuous label space, avoiding unrealistic label flips that violate semantic proximity. Finally, the use of $\text{std}_{\max}$ allows precise control over noise severity while yielding nontrivial and nonuniform noise distributions. Table B.1 summarizes the average MAE of the generated noisy labels over five random seeds under different values of $\text{std}_{\max}$.

Table B.1: Average MAE of generated noisy labels under different noise levels ($\text{std}_{\max}$).

Dataset	$\textbf{std}_{\max} = 30$	$\textbf{std}_{\max} = 60$	$\textbf{std}_{\max} = 90$
UTKFace (Age)	7.04	10.52	11.44
Biwi Kinect (Yaw)	11.11	19.24	24.07
Biwi Kinect (Pitch)	10.98	18.24	22.16
Biwi Kinect (Roll)	9.64	13.61	15.53

Baseline Methods. We compare GRAIN with a diverse set of regression and noisy-label learning baselines, including both methods originally proposed for regression and classical approaches from noisy-label classification adapted to regression settings.

- **Existing Regression Methods.**
  - **Vanilla (MSE).** This baseline directly trains a regression network using the mean squared error (MSE) loss on noisy labels.
  - **SuperLoss** [7]. SuperLoss is a self-paced robust loss function that dynamically reweights training samples according to their loss values, automatically downweighting hard samples with large losses during optimization.
  - **RDI and AUX** [5]. RDI and AUX are robust regression methods designed to mitigate overfitting to mislabeled data. RDI regularizes learning by constraining the distance between the network parameters and their initialization, while AUX introduces a trainable auxiliary variable for each training example to absorb label noise.
  - **C-Mixup** [6]. C-Mixup extends the Mixup strategy to regression by adjusting the sampling probability based on label similarity, encouraging smoother regression functions under noisy supervision.
- **Classification-based noisy-label methods.** We additionally include several classical noisy-label learning methods originally developed for classification, which can be adapted to regression settings:
  - **CDR** [3] divides all parameters into the critical and non-critical ones, and then performs different update rules for different types of parameters.
  - **CoTeaching** [2] jointly trains two networks and discards samples with large losses, based on the observation that clean samples tend to have smaller losses in early training stages.
  - **CNLCU-S and CNLCU-H** [8] identify clean samples via sample selection guided by uncertainty estimation, with CNLCU-H further incorporating historical information.
  - **Selfie** [4] adopts a self-training strategy by iteratively refining pseudo labels using model predictions.
  - **CoSelfie** [4] extends Selfie by combining self-training with co-training to alleviate confirmation bias.

Additionally, for all the methods above, we also implement a cross-entropy loss based version by discretizing the continuous target space into ordered bins and introducing a histogram head, following the same binning strategy used in GRAIN. In all experimental results, these adapted variants are denoted by "BIN". Unless otherwise specified, all baselines share the same network backbone, training protocol, and model selection strategy as GRAIN to ensure a fair comparison.

B.2 Additional Implementation Details

Algorithm 1 GRAIN Training Procedure

1: **Input:** Training data $\mathcal{D} = \{(\mathrm{X}_i, \widetilde{\mathrm{Y}}_i)\}_{i=1}^N$, loss weights $\lambda_{\text{reg}}, \lambda_{\text{ce}}, \lambda_{\text{pos}}, \lambda_{\text{neg}}$, warm-up epochs t_{warm}, total epochs T
2: **Initialize:** Feature extractor θ_{f}, regression head θ_{reg}, histogram head θ_{hist}, residual head θ_{res}; pseudo labels $\{\mathrm{Y}_i^{*(0)} \leftarrow \widetilde{\mathrm{Y}}_i\}_{i=1}^N$; confidences $\{c_i^{*(0)} \leftarrow 1\}_{i=1}^N$
3: **Precompute:** Adaptive bin boundaries and representative values $\{\bar{\mathrm{Y}}_k\}_{k=1}^K$ via (B.1)
4: `// Warm Up`
5: **for** epoch $t = 1$ to t_{warm} **do**
6: Train $(\theta_{\text{f}}, \theta_{\text{reg}}, \theta_{\text{hist}})$ by minimizing $\lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}$
7: Update pseudo labels $\{\mathrm{Y}_i^{*(t)}\}_{i=1}^N$ and confidences $\{c_i^{*(t)}\}_{i=1}^N$ via (B.5) and (B.6) if $t \geq \lfloor t_{\text{warm}}/2 \rfloor$
8: Update $\tau^{(t)}$ via (B.7) if $t \geq \lfloor t_{\text{warm}}/2 \rfloor$
9: **end for**
10: **for** epoch $t = t_{\text{warm}} + 1$ to T **do**
11: `// Pseudo Label and Threshold Update`
12: Update pseudo labels $\{\mathrm{Y}_i^{*(t)}\}_{i=1}^N$ and confidences $\{c_i^{*(t)}\}_{i=1}^N$ via (B.5) and (B.6)
13: Update residual capacity threshold $\tau^{(t)}$ via (B.7)
14: `// Residual Head Training`
15: Freeze $(\theta_{\text{f}}, \theta_{\text{reg}}, \theta_{\text{hist}})$; update θ_{res} by minimizing confidence-weighted residual loss (9)
16: `// Main Model Training`
17: Compute confidence-weighted regression loss (7) and cross-entropy loss (10)
18: Construct contrastive triplets using $\{\mathrm{Y}_i^{*(t)}\}$; compute $\mathcal{R}_{\text{pos}}$ and $\mathcal{R}_{\text{neg}}$ via (11)–(12)
19: Freeze θ_{res}; update $(\theta_{\text{f}}, \theta_{\text{reg}}, \theta_{\text{hist}})$ by minimizing joint loss (13)
20: **end for**
21: **Output:** Trained model parameters $\vartheta = (\theta_{\text{f}}, \theta_{\text{reg}}, \theta_{\text{hist}}, \theta_{\text{res}})$

Binning procedure (Line 3 of Algorithm 1). Before model training, we adaptively discretize the noisy labels according to their empirical distribution, and keep the number of bins fixed throughout the entire training process. Specifically, using Python package `scipy.stats`, we first perform density estimation on the noisy label set $\{\widetilde{Y}_i\}_{i=1}^N$ using a Gaussian kernel density estimator (KDE). To prevent excessively populated bins from dominating the training dynamics, we determine the bin width based on a local quantile range around the peak of the estimated density. Let $\widehat{F}(\cdot)$ denote the cumulative distribution function (CDF) induced by the KDE, and define the corresponding $100q$th quantile function as $\widehat{Q}(q) \triangleq \inf\left\{y \in \mathcal{Y} : \widehat{F}(y) \geq q\right\}$, where $0 < q < 1$. Let y_{peak} denote the location of the density peak and $q_{\text{peak}} \triangleq \widehat{F}(y_{\text{peak}})$ be its associated quantile. We then define the bin width as

$$\texttt{width} \triangleq \widehat{Q}(q_{\text{peak}} + 5\%) - \widehat{Q}(q_{\text{peak}} - 5\%), \tag{B.1}$$

corresponding to a symmetric 10% quantile interval around the peak y_{peak}. This adaptive discretization reflects the underlying label density while mitigating excessive concentration in high-density regions. Given the bin width (B.1), the number of bins is automatically determined as $K \triangleq \lfloor(\mathrm{y}_{\max} - \mathrm{y}_{\min})/\texttt{width}\rfloor$, where $\mathrm{y}_{\max} \triangleq \max_{1\leq i\leq N} \widetilde{\mathrm{Y}}_i$ and $\mathrm{y}_{\min} \triangleq \min_{1\leq i\leq N} \widetilde{\mathrm{Y}}_i$. In the resulting K bins, we let $\bar{\mathrm{y}}_k$ denote a representative value of bin k, which can be chosen as the left endpoint, midpoint, or right endpoint of the corresponding bin interval.

Construction of Robust Pseudo Labels from Histogram Head (Lines 7 and 12 of Algorithm 1). We construct robust pseudo labels from the histogram head. Specifically, in the t-th epoch, let $p_i^{(t)} \in \Delta^{K-1}$ denote the softmax output of the histogram head for input X_i. We define $\mathcal{T}_i^{(t)} \subset [K]$ as the minimal set of highest-probability bins whose cumulative probability mass exceeds 68%:

$$\mathcal{T}_i^{(t)} = \arg\min_{\mathcal{S}\subseteq[K]} |\mathcal{S}| \quad \text{s.t.} \quad \sum_{k\in\mathcal{S}} p_i^{(t)}[k] \geq 0.68, \tag{B.2}$$

where bins are added in decreasing order of $p_i^{(t)}[k]$. We then define the normalized weights as $p_i^{(t)*}[k] = \frac{p_i^{(t)}[k]}{\sum_{k' \in \mathcal{T}_i^{(t)}} p_i^{(t)}[k']}$ for $k \in \mathcal{T}_i^{(t)}$, and $p_i^{(t)*}[k] = 0$ for $k \notin \mathcal{T}_i^{(t)}$. The resulting weighted pseudo label is

$$\breve{\mathrm{Y}}_i^{(t)*} \triangleq \sum_{k=1}^{K} p_i^{(t)*}[k] \cdot \bar{\mathrm{y}}_k, \tag{B.3}$$

where $\bar{\mathrm{y}}_k$ is the representative value of bin k from the binning procedure. Alongside the pseudo label, we define a confidence score that reflects the sharpness of the histogram head's prediction:

$$\breve{c}_i^{(t)} \triangleq 1 - \frac{|\mathcal{T}_i^{(t)}|}{K} \in [0, 1], \tag{B.4}$$

where $|\mathcal{T}_i^{(t)}|$ is the number of selected bins. Intuitively, when the histogram head concentrates its probability mass on fewer bins, the pseudo label is more reliable and $\breve{c}_i^{(t)}$ is closer to 1; conversely, when the probability mass is spread across many bins, the pseudo label is less trustworthy and $\breve{c}_i^{(t)}$ is closer to 0.

To further stabilize training, we maintain Exponential Moving Averages (EMA) of both the pseudo label and confidence score across epochs. The pseudo labels and confidence scores are initialized as $\mathrm{Y}_i^{*(0)} = \widetilde{\mathrm{Y}}_i$ and $c_i^{*(0)} = 1$, respectively, and remain unchanged for $t < \lfloor t_{\text{warm}}/2 \rfloor$. Starting from epoch $t = \lfloor t_{\text{warm}}/2 \rfloor$, both are updated together at each epoch, including during the remaining warm-up epochs, using the following EMA rules:

$$\mathrm{Y}_i^{(t)*} \leftarrow \alpha\, \mathrm{Y}_i^{(t-1)*} + (1 - \alpha)\, \breve{\mathrm{Y}}_i^{(t)*}, \tag{B.5}$$

$$c_i^{(t)*} \leftarrow \alpha\, c_i^{(t-1)*} + (1 - \alpha)\, \breve{c}_i^{(t)}, \tag{B.6}$$

where $\alpha \in (0, 1)$ is the EMA decay factor (we use $\alpha = 0.9$ in our experiments). During training, the per-instance regression and residual losses are weighted by their corresponding pseudo-label confidence c_i^* before averaging across training instances. This construction achieves robustness by concentrating supervision on the most confident bins via the 68% cumulative threshold, with the confidence score naturally down-weighting pseudo labels when the histogram head spreads probability mass across many bins, and the EMA updates smoothing out per-epoch fluctuations for more stable training.

Residual Capacity τ Update (Lines 8 and 13 of Algorithm 1). Starting at epoch $t = \lfloor t_{\text{warm}}/2 \rfloor$, we update the residual head capacity τ based on the residual Median Absolute Deviation (MAD) [17]. Let $r_i^{(t)} = |\widetilde{\mathrm{Y}}_i - \mathrm{Y}_i^{(t)*}|$ denote the absolute residual between the noisy label and the current pseudo label for sample i. We compute the robust scale estimate

$$\breve{\tau}^{(t)} \triangleq 1.345 \cdot \text{median}\left(\left\{ \left| r_i^{(t)} - \text{median}\left(\{r_j^{(t)}\}_{j=1}^{N} \right) \right| \right\}_{i=1}^{N} \right),$$

which is a robust and outlier-resistant estimator of the noise scale. The EMA update then follows

$$\tau^{(t)} \leftarrow \alpha\, \tau^{(t-1)} + (1 - \alpha)\, \breve{\tau}^{(t)}, \tag{B.7}$$

with the same decay factor α as in (B.5)–(B.6). This update begins at epoch $t = \lfloor t_{\text{warm}}/2 \rfloor$, after the pseudo labels and confidence scores have been updated, at which point τ is initialized to $\breve{\tau}^{(t)}$.

Hyperparameter Selection. The implementation of GRAIN involves several hyperparameters.

(1) **Loss weights for the regression and histogram heads.** Observing that the magnitude of the cross-entropy loss varies less than that of the regression loss during training, we fix the classification loss weight $\lambda_{\text{ce}} = 1.0$, and gradually ramp up the regression loss weight λ_{reg} from 0 to its maximum value over the first 20 epochs, with the maximum value taken from $\{0.2, 0.5\}$.

(2) **Residual head training.** As in the regression head, the parameter τ is automatically determined during training and does not require manual tuning.

(3) **Across-bin contrastive regularization loss.** For the bin-based contrastive regularization, we select λ_{pos} from $\{0.01, 0.05\}$. Instead of tuning λ_{neg} independently, we fix the ratio $\lambda_{\text{neg}}/\lambda_{\text{pos}} = 0.5$ throughout all experiments.

Computational Infrastructure. All experiments were conducted on a single NVIDIA L40S GPU with 48GB memory. The average training time per run of GRAIN (across five random seeds) is approximately 2.5h on UTKFace, 2.3h on Biwi Kinect, 3.5h on Bone-Age, and 0.8h on APPA-REAL. The additional components introduced by GRAIN (the residual head, histogram head, and bin-level contrastive objective) incur negligible overhead relative to the backbone training cost.

B.3 Extended Results

Hyperparameter Sensitivity Analysis. Figures B.1 and B.2 present the hyperparameter sensitivity analysis for λ_{pos} and the ratio $\lambda_{\text{neg}}/\lambda_{\text{pos}}$, respectively, evaluated on the UTKFace dataset under varying levels of simulated Gaussian label corruption. In Figure B.1, the red dashed line indicates the performance without positive contrastive regularization ($\lambda_{\text{pos}} = 0$), while in Figure B.2, it indicates the performance without negative contrastive regularization ($\lambda_{\text{neg}} = 0$). The black dashed lines denote the best-performing baseline method. Overall, GRAIN remains robust to the choice of contrastive regularization coefficients within a reasonable range across different noise levels. Figure B.3 shows that performance remains stable across different numbers of contrastive pairs. Figures B.4 and B.5 show that performance remains stable across different numbers of bins.

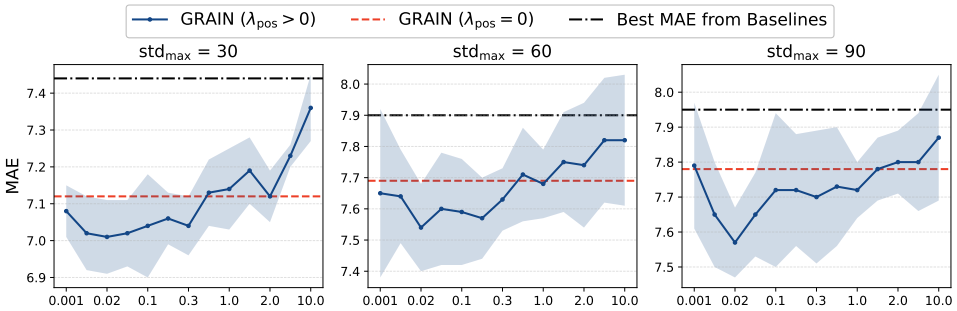


Figure B.1: Hyperparameter sensitivity analysis for λ_{pos} on UTKFace.

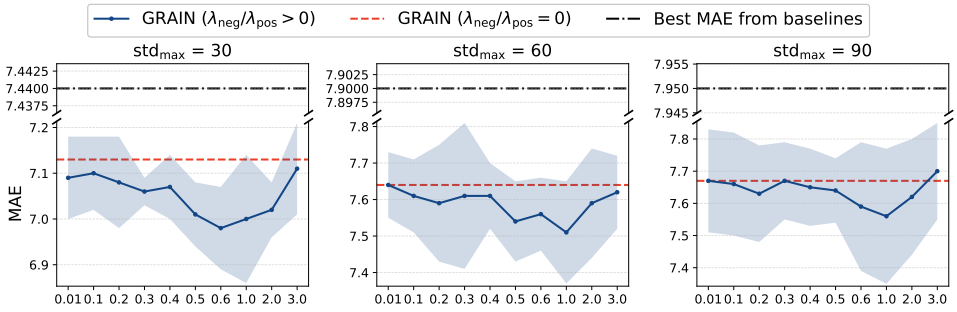


Figure B.2: Hyperparameter sensitivity analysis for $\lambda_{\text{neg}}/\lambda_{\text{pos}}$ on UTKFace.

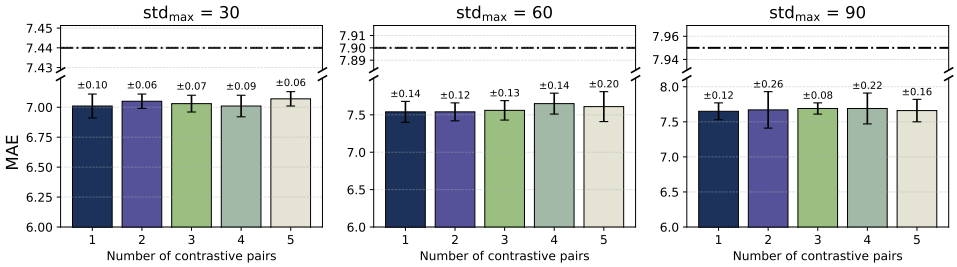


Figure B.3: Hyperparameter sensitivity analysis on the number of contrastive pairs on UTKFace. Black dashed lines denote the best-performing baseline method.

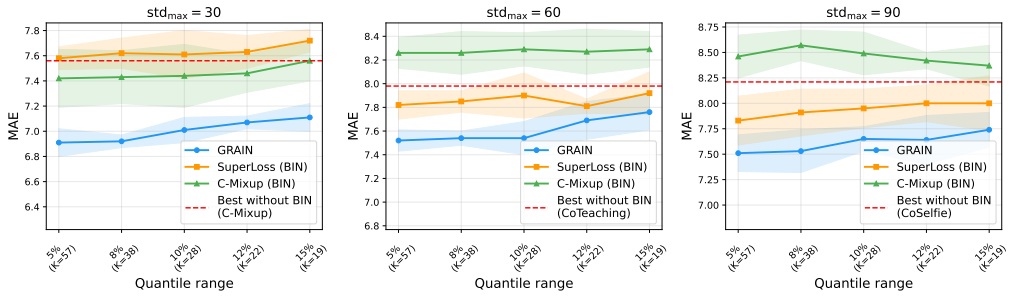


Figure B.4: Sensitivity analysis of the binning strategy on UTKFace. The x-axis shows the symmetric quantile interval around the density peak y_{peak} used to determine the bin width (Equation (B.1)), with the corresponding number of bins K indicated in parentheses. A larger quantile interval yields a wider bin width and thus fewer bins. GRAIN (■) is compared against two strong BIN-enhanced baselines, C-Mixup (BIN) (■) and SuperLoss (BIN) (■). The red dashed line (- - -) indicates the best-performing baseline without binning. Shaded regions indicate ± 1 standard deviation across 5 random seeds.

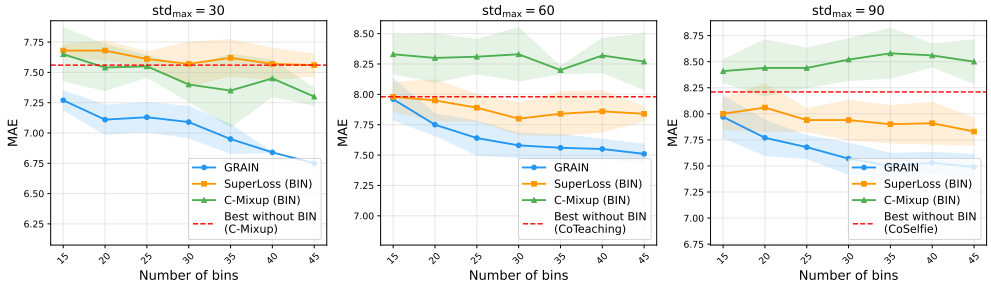


Figure B.5: Sensitivity analysis of the binning strategy on UTKFace. The x-axis shows the number of bins. GRAIN (■) is compared against two strong BIN-enhanced baselines, C-Mixup (BIN) (■) and SuperLoss (BIN) (■). The red dashed line (- - -) indicates the best-performing baseline without binning. Shaded regions indicate ± 1 standard deviation across 5 random seeds.

UOT, OT, and Cosine Similarity. Figure B.6 compares the performance of GRAIN when using UOT, balanced OT, and cosine similarity as the discrepancy measure in the across-bin contrastive feature regularization.

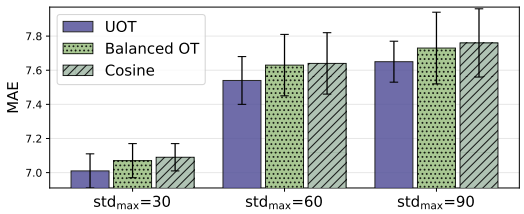


Figure B.6: MAE comparison on UTKFace when replacing UOT with balanced OT or cosine similarity in the across-bin contrastive feature regularization. Error bars indicate ± 1 standard deviation across 5 random seeds.

Residual Analysis. The scatter plot in Figure B.7 illustrates that the residual head captures label noise rather than predictive signals.

τ-Dynamics. Figure B.8 shows the dynamic evolution of the adaptive parameter τ throughout training

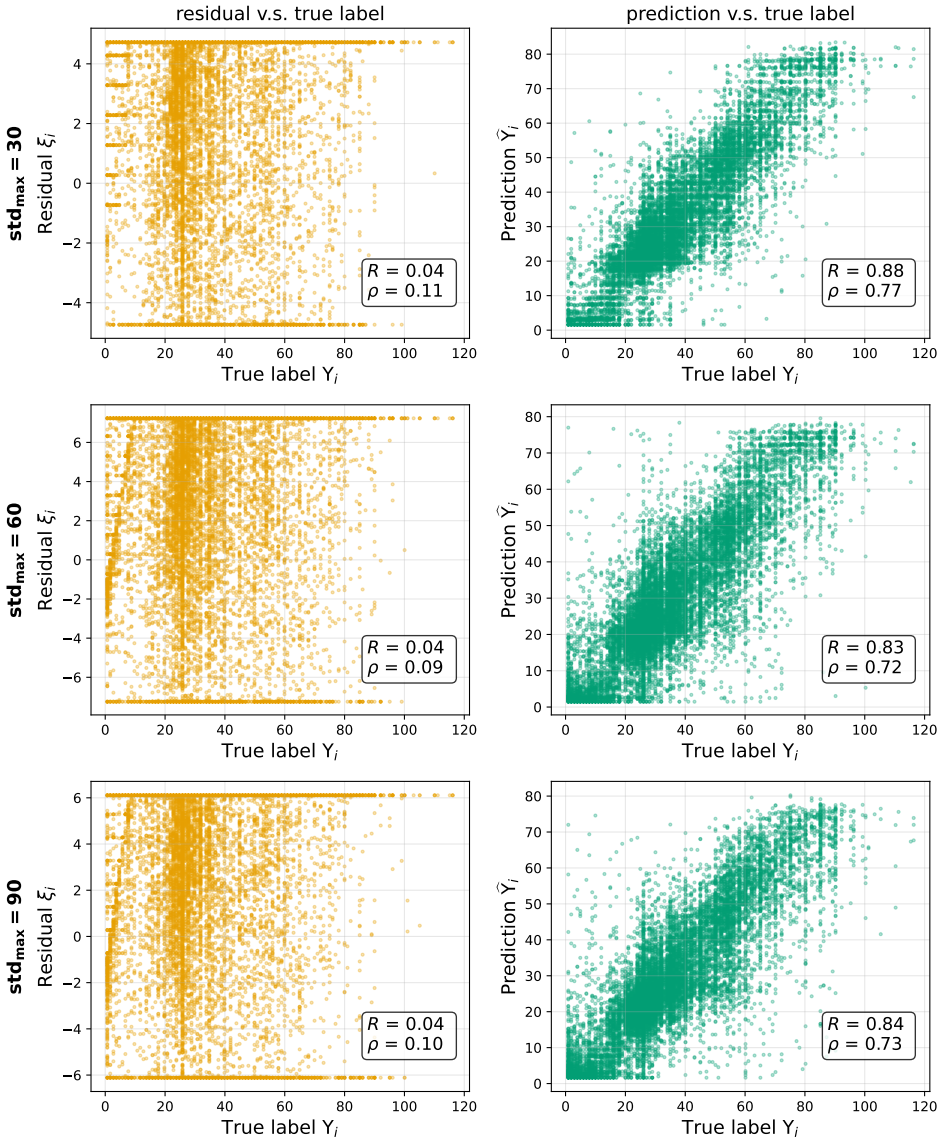


Figure B.7: Scatter plots of residual ξ_i vs. true label Y_i (left column) and prediction $\widehat{\mathrm{Y}}_i$ vs. true label Y_i (right column) on UTKFace, evaluated at the last training epoch using seed 1. Both Pearson R and Spearman ρ are reported in each panel.

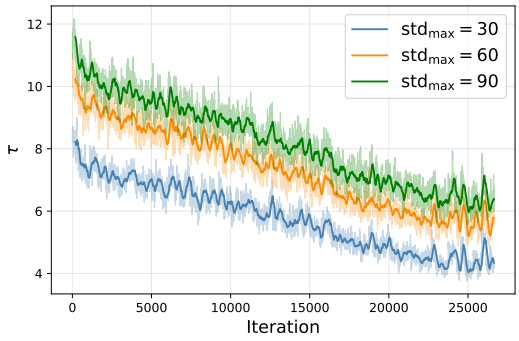


Figure B.8: Dynamic evolution of the adaptive parameter τ throughout training on UTKFace. The light curves show the raw τ values, while the dark curves show their smoothed trajectories via moving average.

Results on APPA-REAL. Results on APPA-REAL with real-world label noise are reported in Table B.2, demonstrating the effectiveness of GRAIN in handling regression label noise across different tasks.

Table B.2: Mean ± standard deviation on APPA-REAL dataset with real-world label noise. The best baseline per metric is highlighted in blue, and the results produced from our proposed GRAIN and GRAIN-AUG methods are presented in **boldface**.

Method	APPA-REAL		
	MAE ↓	**RMSE** ↓	R^2 ↑
Vanilla	$10.75_{\pm 0.14}$	$14.83_{\pm 0.16}$	$0.40_{\pm 0.01}$
SuperLoss	$12.67_{\pm 0.14}$	$16.77_{\pm 0.18}$	$0.24_{\pm 0.02}$
RDI	$10.71_{\pm 0.07}$	$14.80_{\pm 0.13}$	$0.41_{\pm 0.01}$
AUX	$10.72_{\pm 0.16}$	$14.81_{\pm 0.18}$	$0.41_{\pm 0.01}$
C-Mixup	$10.66_{\pm 0.08}$	$14.73_{\pm 0.14}$	$0.41_{\pm 0.01}$
CDR	$10.73_{\pm 0.16}$	$14.81_{\pm 0.22}$	$0.41_{\pm 0.02}$
Selfie	$12.78_{\pm 0.09}$	$16.95_{\pm 0.12}$	$0.22_{\pm 0.01}$
CoTeaching †	$12.67_{\pm 0.11}$	$16.85_{\pm 0.16}$	$0.23_{\pm 0.01}$
CNLCU-S †	$12.76_{\pm 0.09}$	$16.95_{\pm 0.11}$	$0.22_{\pm 0.01}$
CNLCU-H †	$12.73_{\pm 0.10}$	$16.96_{\pm 0.25}$	$0.22_{\pm 0.02}$
CoSelfie †	$12.78_{\pm 0.09}$	$16.95_{\pm 0.12}$	$0.22_{\pm 0.01}$
Vanilla (BIN)	$10.77_{\pm 0.29}$	$15.69_{\pm 0.48}$	$0.33_{\pm 0.04}$
SuperLoss (BIN)	$11.41_{\pm 0.15}$	$16.03_{\pm 0.18}$	$0.30_{\pm 0.02}$
RDI (BIN)	$10.46_{\pm 0.09}$	$14.78_{\pm 0.06}$	$0.41_{\pm 0.01}$
AUX (BIN)	$10.41_{\pm 0.13}$	$14.76_{\pm 0.23}$	$0.41_{\pm 0.02}$
C-Mixup (BIN)	$10.43_{\pm 0.08}$	$14.98_{\pm 0.19}$	$0.39_{\pm 0.02}$
CDR (BIN)	$10.69_{\pm 0.10}$	$15.62_{\pm 0.28}$	$0.34_{\pm 0.02}$
Selfie (BIN)	$11.69_{\pm 0.22}$	$16.79_{\pm 0.22}$	$0.24_{\pm 0.02}$
CoTeaching (BIN) †	$11.84_{\pm 0.22}$	$16.92_{\pm 0.31}$	$0.22_{\pm 0.03}$
CNLCU-S (BIN) †	$12.21_{\pm 0.48}$	$17.52_{\pm 0.86}$	$0.17_{\pm 0.08}$
CNLCU-H (BIN) †	$12.09_{\pm 0.27}$	$17.34_{\pm 0.46}$	$0.18_{\pm 0.04}$
CoSelfie (BIN) †	$11.69_{\pm 0.22}$	$16.79_{\pm 0.22}$	$0.24_{\pm 0.02}$
GRAIN (ours)	$\mathbf{10.22_{\pm 0.15}}$	$\mathbf{14.66_{\pm 0.21}}$	$\mathbf{0.42_{\pm 0.02}}$
GRAIN-AUG (ours)	$\mathbf{10.00_{\pm 0.12}}$	$\mathbf{14.48_{\pm 0.16}}$	$\mathbf{0.43_{\pm 0.01}}$

Results on Biwi Kinect. Tables B.3, B.4, and B.5 present the Biwi Kinect results for yaw, pitch, and roll estimation under synthetic Gaussian label noise across varying corruption levels, evaluated with the three metrics described in Appendix B.1. GRAIN and GRAIN-AUG consistently outperform all baselines across nearly all noise levels and metrics, demonstrating strong robustness to synthetic label noise.

Table B.3: Mean ± standard deviation on Biwi Kinect (Yaw) under simulated Gaussian label noise with corruption levels controlled by $\text{std}_{\max}$. The best baseline per metric is highlighted in blue, and the results produced from our proposed GRAIN and GRAIN-AUG methods are presented in **boldface**.

Method	$\text{std}_{\max}$= 30			$\text{std}_{\max}$= 60			$\text{std}_{\max}$= 90		
	MAE ↓	**RMSE** ↓	R^2 ↑	**MAE** ↓	**RMSE** ↓	R^2 ↑	**MAE** ↓	**RMSE** ↓	R^2 ↑
Vanilla	$5.33_{\pm 0.85}$	$7.66_{\pm 1.65}$	$0.91_{\pm 0.03}$	$7.54_{\pm 1.09}$	$10.61_{\pm 1.30}$	$0.84_{\pm 0.03}$	$9.79_{\pm 1.23}$	$13.93_{\pm 1.87}$	$0.73_{\pm 0.05}$
SuperLoss	$4.79_{\pm 0.93}$	$7.26_{\pm 2.28}$	$0.92_{\pm 0.04}$	$6.09_{\pm 0.66}$	$8.46_{\pm 1.19}$	$0.90_{\pm 0.03}$	$6.70_{\pm 1.43}$	$9.61_{\pm 2.56}$	$0.86_{\pm 0.06}$
RDI	$5.02_{\pm 0.87}$	$7.19_{\pm 2.15}$	$0.92_{\pm 0.04}$	$7.24_{\pm 1.08}$	$9.97_{\pm 1.75}$	$0.86_{\pm 0.03}$	$9.67_{\pm 1.17}$	$13.63_{\pm 1.53}$	$0.74_{\pm 0.03}$
AUX	$5.87_{\pm 1.07}$	$8.56_{\pm 2.24}$	$0.89_{\pm 0.05}$	$7.22_{\pm 1.13}$	$9.99_{\pm 1.75}$	$0.86_{\pm 0.03}$	$9.60_{\pm 0.67}$	$13.29_{\pm 0.96}$	$0.75_{\pm 0.04}$
C-Mixup	$4.75_{\pm 0.70}$	$6.94_{\pm 1.38}$	$0.93_{\pm 0.03}$	$6.49_{\pm 1.04}$	$9.84_{\pm 2.13}$	$0.86_{\pm 0.05}$	$8.71_{\pm 0.97}$	$11.75_{\pm 0.93}$	$0.80_{\pm 0.05}$
CDR	$5.26_{\pm 0.87}$	$7.78_{\pm 1.89}$	$0.91_{\pm 0.03}$	$7.66_{\pm 1.29}$	$10.65_{\pm 1.75}$	$0.84_{\pm 0.03}$	$9.46_{\pm 0.50}$	$13.29_{\pm 0.74}$	$0.73_{\pm 0.04}$
Selfie	$5.22_{\pm 1.03}$	$7.99_{\pm 2.45}$	$0.90_{\pm 0.05}$	$6.32_{\pm 0.94}$	$9.08_{\pm 2.10}$	$0.88_{\pm 0.05}$	$7.44_{\pm 1.26}$	$10.92_{\pm 2.84}$	$0.81_{\pm 0.09}$
CoTeaching †	$4.93_{\pm 0.75}$	$7.71_{\pm 1.95}$	$0.91_{\pm 0.04}$	$5.84_{\pm 1.28}$	$8.64_{\pm 2.58}$	$0.89_{\pm 0.05}$	$6.68_{\pm 0.47}$	$9.67_{\pm 1.66}$	$0.86_{\pm 0.05}$
CNLCU-S †	$4.97_{\pm 0.86}$	$7.38_{\pm 1.81}$	$0.92_{\pm 0.03}$	$6.11_{\pm 1.00}$	$8.75_{\pm 1.83}$	$0.89_{\pm 0.04}$	$6.42_{\pm 0.62}$	$9.26_{\pm 1.68}$	$0.87_{\pm 0.05}$
CNLCU-H †	$4.88_{\pm 0.86}$	$7.27_{\pm 1.84}$	$0.92_{\pm 0.03}$	$6.09_{\pm 1.27}$	$8.86_{\pm 2.27}$	$0.88_{\pm 0.05}$	$6.41_{\pm 0.78}$	$9.42_{\pm 1.96}$	$0.87_{\pm 0.05}$
CoSelfie †	$5.08_{\pm 1.00}$	$7.81_{\pm 2.63}$	$0.91_{\pm 0.05}$	$5.94_{\pm 1.17}$	$8.76_{\pm 2.16}$	$0.88_{\pm 0.05}$	$6.74_{\pm 1.16}$	$10.08_{\pm 2.23}$	$0.84_{\pm 0.09}$
Vanilla (BIN)	$6.18_{\pm 1.44}$	$9.03_{\pm 2.88}$	$0.88_{\pm 0.06}$	$7.33_{\pm 1.76}$	$10.14_{\pm 2.74}$	$0.85_{\pm 0.07}$	$7.75_{\pm 1.28}$	$11.08_{\pm 2.46}$	$0.82_{\pm 0.07}$
SuperLoss (BIN)	$5.55_{\pm 1.36}$	$8.09_{\pm 2.16}$	$0.90_{\pm 0.04}$	$6.62_{\pm 1.21}$	$9.48_{\pm 2.52}$	$0.87_{\pm 0.06}$	$7.17_{\pm 0.94}$	$10.09_{\pm 1.92}$	$0.85_{\pm 0.05}$
RDI (BIN)	$5.76_{\pm 1.09}$	$8.29_{\pm 2.07}$	$0.90_{\pm 0.04}$	$6.91_{\pm 1.45}$	$9.45_{\pm 2.40}$	$0.87_{\pm 0.05}$	$7.66_{\pm 1.04}$	$10.81_{\pm 1.89}$	$0.82_{\pm 0.06}$
AUX (BIN)	$5.11_{\pm 0.92}$	$7.59_{\pm 2.04}$	$0.91_{\pm 0.04}$	$6.91_{\pm 1.45}$	$9.45_{\pm 2.40}$	$0.87_{\pm 0.05}$	$7.30_{\pm 1.17}$	$10.18_{\pm 2.12}$	$0.85_{\pm 0.05}$
C-Mixup (BIN)	$5.61_{\pm 1.12}$	$8.26_{\pm 2.38}$	$0.90_{\pm 0.05}$	$7.21_{\pm 1.73}$	$10.33_{\pm 2.86}$	$0.84_{\pm 0.07}$	$7.41_{\pm 1.68}$	$10.10_{\pm 2.27}$	$0.85_{\pm 0.06}$
CDR (BIN)	$5.99_{\pm 1.03}$	$8.43_{\pm 1.63}$	$0.90_{\pm 0.03}$	$7.45_{\pm 1.41}$	$10.47_{\pm 2.50}$	$0.83_{\pm 0.07}$	$7.57_{\pm 1.71}$	$10.77_{\pm 2.59}$	$0.83_{\pm 0.07}$
Selfie (BIN)	$6.82_{\pm 1.55}$	$9.92_{\pm 2.83}$	$0.85_{\pm 0.07}$	$6.49_{\pm 1.06}$	$9.17_{\pm 2.07}$	$0.88_{\pm 0.05}$	$8.10_{\pm 1.19}$	$11.23_{\pm 2.00}$	$0.81_{\pm 0.07}$
CoTeaching (BIN) †	$6.08_{\pm 1.25}$	$8.90_{\pm 2.14}$	$0.88_{\pm 0.05}$	$7.49_{\pm 1.14}$	$10.52_{\pm 2.02}$	$0.84_{\pm 0.05}$	$9.92_{\pm 2.44}$	$14.79_{\pm 4.51}$	$0.70_{\pm 0.11}$
CNLCU-S (BIN) †	$5.46_{\pm 1.14}$	$8.12_{\pm 2.31}$	$0.90_{\pm 0.05}$	$7.27_{\pm 1.38}$	$10.01_{\pm 1.77}$	$0.85_{\pm 0.04}$	$8.48_{\pm 2.07}$	$12.52_{\pm 3.77}$	$0.74_{\pm 0.19}$
CNLCU-H (BIN) †	$6.19_{\pm 1.08}$	$8.67_{\pm 1.84}$	$0.89_{\pm 0.04}$	$7.35_{\pm 1.60}$	$10.34_{\pm 2.89}$	$0.84_{\pm 0.07}$	$7.73_{\pm 1.00}$	$10.96_{\pm 1.60}$	$0.83_{\pm 0.04}$
CoSelfie (BIN) †	$6.26_{\pm 1.67}$	$8.94_{\pm 3.18}$	$0.88_{\pm 0.07}$	$8.15_{\pm 1.54}$	$11.51_{\pm 2.33}$	$0.80_{\pm 0.07}$	$9.21_{\pm 1.08}$	$13.10_{\pm 2.02}$	$0.75_{\pm 0.07}$
GRAIN (ours)	$\mathbf{4.88_{\pm 0.86}}$	$\mathbf{6.49_{\pm 0.67}}$	$\mathbf{0.93_{\pm 0.03}}$	$\mathbf{5.57_{\pm 0.89}}$	$\mathbf{7.87_{\pm 1.38}}$	$\mathbf{0.90_{\pm 0.04}}$	$\mathbf{6.00_{\pm 0.88}}$	$\mathbf{8.46_{\pm 1.58}}$	$\mathbf{0.88_{\pm 0.07}}$
GRAIN-AUG (ours)	$\mathbf{4.76_{\pm 0.47}}$	$\mathbf{6.59_{\pm 0.12}}$	$\mathbf{0.93_{\pm 0.02}}$	$\mathbf{5.30_{\pm 0.61}}$	$\mathbf{7.18_{\pm 0.89}}$	$\mathbf{0.92_{\pm 0.03}}$	$\mathbf{5.77_{\pm 0.81}}$	$\mathbf{8.03_{\pm 1.16}}$	$\mathbf{0.90_{\pm 0.03}}$

Table B.4: Mean ± standard deviation on Biwi Kinect (Pitch) under simulated Gaussian label noise with corruption levels controlled by $\text{std}_{\max}$. The best baseline per metric is highlighted in blue, and the results produced from our proposed GRAIN and GRAIN-AUG methods are presented in **boldface**.

Method	$\text{std}_{\max}$= 30			$\text{std}_{\max}$= 60			$\text{std}_{\max}$= 90		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
Vanilla	$6.18_{\pm 1.78}$	$8.77_{\pm 3.01}$	$0.85_{\pm 0.11}$	$8.01_{\pm 1.97}$	$11.48_{\pm 2.71}$	$0.77_{\pm 0.10}$	$10.59_{\pm 1.83}$	$14.93_{\pm 2.24}$	$0.62_{\pm 0.12}$
SuperLoss	$5.78_{\pm 1.36}$	$8.07_{\pm 2.18}$	$0.88_{\pm 0.06}$	$6.80_{\pm 1.60}$	$9.43_{\pm 2.05}$	$0.85_{\pm 0.06}$	$8.12_{\pm 2.01}$	$11.75_{\pm 3.10}$	$0.76_{\pm 0.12}$
RDI	$6.25_{\pm 1.68}$	$8.81_{\pm 2.55}$	$0.86_{\pm 0.07}$	$8.15_{\pm 1.81}$	$11.67_{\pm 2.81}$	$0.76_{\pm 0.12}$	$10.10_{\pm 1.75}$	$14.36_{\pm 2.65}$	$0.65_{\pm 0.13}$
AUX	$6.17_{\pm 1.76}$	$8.61_{\pm 2.42}$	$0.86_{\pm 0.08}$	$8.23_{\pm 1.66}$	$11.85_{\pm 2.88}$	$0.75_{\pm 0.12}$	$10.05_{\pm 1.55}$	$13.77_{\pm 2.07}$	$0.68_{\pm 0.10}$
C-Mixup	$5.74_{\pm 1.71}$	$8.21_{\pm 2.78}$	$0.87_{\pm 0.08}$	$7.36_{\pm 1.25}$	$10.15_{\pm 1.86}$	$0.83_{\pm 0.06}$	$9.35_{\pm 2.04}$	$13.21_{\pm 2.71}$	$0.69_{\pm 0.13}$
CDR	$6.61_{\pm 1.87}$	$9.17_{\pm 2.55}$	$0.85_{\pm 0.08}$	$8.36_{\pm 1.61}$	$12.02_{\pm 2.83}$	$0.74_{\pm 0.14}$	$10.38_{\pm 1.78}$	$14.35_{\pm 2.26}$	$0.65_{\pm 0.11}$
Selfie	$5.98_{\pm 1.82}$	$8.68_{\pm 3.11}$	$0.86_{\pm 0.10}$	$8.38_{\pm 2.20}$	$12.20_{\pm 2.84}$	$0.74_{\pm 0.13}$	$11.06_{\pm 2.95}$	$15.13_{\pm 3.24}$	$0.60_{\pm 0.18}$
CoTeaching †	$6.13_{\pm 1.95}$	$8.53_{\pm 2.95}$	$0.86_{\pm 0.08}$	$7.59_{\pm 2.66}$	$11.29_{\pm 3.91}$	$0.76_{\pm 0.15}$	$9.02_{\pm 3.14}$	$13.23_{\pm 4.23}$	$0.68_{\pm 0.18}$
CNLCU-S †	$6.06_{\pm 1.87}$	$8.76_{\pm 2.83}$	$0.86_{\pm 0.08}$	$7.76_{\pm 2.46}$	$11.34_{\pm 3.46}$	$0.77_{\pm 0.13}$	$9.84_{\pm 2.62}$	$14.63_{\pm 3.52}$	$0.62_{\pm 0.17}$
CNLCU-H †	$6.18_{\pm 1.87}$	$8.62_{\pm 2.79}$	$0.86_{\pm 0.08}$	$7.87_{\pm 2.74}$	$11.50_{\pm 3.59}$	$0.76_{\pm 0.15}$	$10.16_{\pm 2.67}$	$15.26_{\pm 3.34}$	$0.60_{\pm 0.17}$
CoSelfie †	$6.07_{\pm 1.82}$	$8.29_{\pm 2.66}$	$0.87_{\pm 0.07}$	$8.02_{\pm 2.51}$	$11.75_{\pm 2.99}$	$0.76_{\pm 0.13}$	$10.24_{\pm 2.61}$	$15.03_{\pm 3.40}$	$0.60_{\pm 0.21}$
Vanilla (BIN)	$6.16_{\pm 1.40}$	$8.91_{\pm 2.59}$	$0.85_{\pm 0.09}$	$7.21_{\pm 1.46}$	$9.42_{\pm 1.95}$	$0.85_{\pm 0.07}$	$8.95_{\pm 2.33}$	$12.92_{\pm 4.10}$	$0.69_{\pm 0.19}$
SuperLoss (BIN)	$5.51_{\pm 1.48}$	$7.53_{\pm 2.12}$	$0.90_{\pm 0.05}$	$7.02_{\pm 0.62}$	$9.21_{\pm 0.82}$	$0.86_{\pm 0.03}$	$8.44_{\pm 2.09}$	$12.52_{\pm 3.36}$	$0.72_{\pm 0.14}$
RDI (BIN)	$6.25_{\pm 1.68}$	$8.81_{\pm 2.55}$	$0.86_{\pm 0.07}$	$7.41_{\pm 1.36}$	$9.69_{\pm 1.78}$	$0.84_{\pm 0.05}$	$8.71_{\pm 2.04}$	$11.98_{\pm 3.14}$	$0.75_{\pm 0.12}$
AUX (BIN)	$6.04_{\pm 1.22}$	$8.22_{\pm 1.91}$	$0.88_{\pm 0.06}$	$7.52_{\pm 1.89}$	$10.11_{\pm 2.80}$	$0.81_{\pm 0.11}$	$8.38_{\pm 2.04}$	$11.58_{\pm 3.28}$	$0.76_{\pm 0.13}$
C-Mixup (BIN)	$5.62_{\pm 0.99}$	$7.75_{\pm 1.68}$	$0.90_{\pm 0.04}$	$6.93_{\pm 1.08}$	$9.57_{\pm 1.63}$	$0.84_{\pm 0.06}$	$8.64_{\pm 2.39}$	$12.22_{\pm 4.35}$	$0.71_{\pm 0.22}$
CDR (BIN)	$5.66_{\pm 1.52}$	$7.54_{\pm 1.92}$	$0.90_{\pm 0.05}$	$8.36_{\pm 1.61}$	$12.02_{\pm 2.83}$	$0.74_{\pm 0.14}$	$10.38_{\pm 1.78}$	$14.35_{\pm 2.26}$	$0.65_{\pm 0.11}$
Selfie (BIN)	$7.44_{\pm 2.62}$	$10.43_{\pm 3.60}$	$0.80_{\pm 0.11}$	$12.81_{\pm 1.97}$	$17.05_{\pm 2.56}$	$0.51_{\pm 0.16}$	$14.83_{\pm 2.38}$	$20.26_{\pm 2.71}$	$0.32_{\pm 0.15}$
CoTeaching (BIN) †	$7.91_{\pm 2.24}$	$11.81_{\pm 3.00}$	$0.75_{\pm 0.13}$	$10.75_{\pm 2.87}$	$15.75_{\pm 4.05}$	$0.58_{\pm 0.18}$	$13.06_{\pm 2.03}$	$17.71_{\pm 3.02}$	$0.47_{\pm 0.19}$
CNLCU-S (BIN) †	$6.78_{\pm 2.11}$	$9.48_{\pm 2.65}$	$0.84_{\pm 0.08}$	$9.64_{\pm 3.18}$	$14.05_{\pm 4.52}$	$0.64_{\pm 0.20}$	$13.35_{\pm 3.06}$	$17.91_{\pm 3.78}$	$0.46_{\pm 0.19}$
CNLCU-H (BIN) †	$7.75_{\pm 2.16}$	$11.35_{\pm 3.47}$	$0.76_{\pm 0.14}$	$9.67_{\pm 2.28}$	$14.38_{\pm 3.24}$	$0.65_{\pm 0.14}$	$13.04_{\pm 2.04}$	$17.89_{\pm 2.77}$	$0.46_{\pm 0.17}$
CoSelfie (BIN) †	$7.25_{\pm 2.44}$	$10.78_{\pm 4.29}$	$0.78_{\pm 0.15}$	$9.29_{\pm 2.51}$	$13.75_{\pm 3.54}$	$0.67_{\pm 0.15}$	$14.08_{\pm 3.40}$	$19.25_{\pm 4.65}$	$0.36_{\pm 0.25}$
GRAIN (ours)	$\mathbf{5.26_{\pm 0.95}}$	$\mathbf{7.37_{\pm 1.66}}$	$\mathbf{0.90_{\pm 0.05}}$	$\mathbf{6.03_{\pm 1.27}}$	$\mathbf{8.12_{\pm 2.00}}$	$\mathbf{0.89_{\pm 0.05}}$	$\mathbf{7.59_{\pm 2.21}}$	$\mathbf{11.05_{\pm 3.60}}$	$\mathbf{0.78_{\pm 0.12}}$
GRAIN-AUG (ours)	$\mathbf{4.97_{\pm 1.41}}$	$\mathbf{6.74_{\pm 1.71}}$	$\mathbf{0.92_{\pm 0.04}}$	$\mathbf{5.83_{\pm 1.75}}$	$\mathbf{8.39_{\pm 3.04}}$	$\mathbf{0.86_{\pm 0.10}}$	$\mathbf{6.88_{\pm 2.36}}$	$\mathbf{9.42_{\pm 3.40}}$	$\mathbf{0.84_{\pm 0.09}}$

Table B.5: Mean ± standard deviation on Biwi Kinect (Roll) under simulated Gaussian label noise with corruption levels controlled by $\text{std}_{\max}$. The best baseline per metric is highlighted in blue, and the results produced from our proposed GRAIN and GRAIN-AUG methods are presented in **boldface**.

Method	$\text{std}_{\max}$= 30			$\text{std}_{\max}$= 60			$\text{std}_{\max}$= 90		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
Vanilla	$5.94_{\pm 2.05}$	$8.78_{\pm 3.34}$	$0.54_{\pm 0.07}$	$7.04_{\pm 2.39}$	$10.11_{\pm 3.59}$	$0.38_{\pm 0.06}$	$8.13_{\pm 2.92}$	$11.95_{\pm 4.62}$	$0.15_{\pm 0.12}$
SuperLoss	$5.70_{\pm 2.19}$	$8.59_{\pm 3.45}$	$0.57_{\pm 0.07}$	$6.98_{\pm 2.13}$	$10.51_{\pm 3.67}$	$0.14_{\pm 0.15}$	$8.17_{\pm 3.07}$	$12.35_{\pm 4.96}$	$0.04_{\pm 0.05}$
RDI	$5.45_{\pm 2.15}$	$7.97_{\pm 3.21}$	$0.63_{\pm 0.06}$	$6.41_{\pm 2.20}$	$9.60_{\pm 3.41}$	$0.44_{\pm 0.07}$	$7.94_{\pm 2.76}$	$11.70_{\pm 4.25}$	$0.17_{\pm 0.15}$
AUX	$5.29_{\pm 2.01}$	$8.04_{\pm 3.17}$	$0.61_{\pm 0.05}$	$6.67_{\pm 2.02}$	$10.02_{\pm 3.34}$	$0.38_{\pm 0.02}$	$7.55_{\pm 2.59}$	$11.25_{\pm 3.85}$	$0.22_{\pm 0.15}$
C-Mixup	$5.29_{\pm 1.88}$	$7.74_{\pm 2.77}$	$0.63_{\pm 0.04}$	$6.80_{\pm 2.49}$	$10.10_{\pm 3.88}$	$0.38_{\pm 0.12}$	$7.85_{\pm 2.93}$	$11.60_{\pm 4.49}$	$0.20_{\pm 0.16}$
CDR	$5.39_{\pm 1.88}$	$8.04_{\pm 3.07}$	$0.61_{\pm 0.07}$	$6.68_{\pm 2.54}$	$9.84_{\pm 3.73}$	$0.32_{\pm 0.06}$	$7.98_{\pm 3.54}$	$11.62_{\pm 5.15}$	$0.24_{\pm 0.08}$
Selfie	$6.73_{\pm 3.11}$	$9.91_{\pm 4.89}$	$0.43_{\pm 0.22}$	$7.52_{\pm 2.21}$	$11.27_{\pm 3.54}$	$0.01_{\pm 0.03}$	$8.40_{\pm 2.87}$	$12.69_{\pm 4.38}$	$0.01_{\pm 0.02}$
CoTeaching †	$6.08_{\pm 2.71}$	$9.19_{\pm 4.26}$	$0.51_{\pm 0.15}$	$8.20_{\pm 2.90}$	$12.26_{\pm 4.54}$	$0.09_{\pm 0.12}$	$8.32_{\pm 2.82}$	$12.45_{\pm 4.47}$	$0.06_{\pm 0.08}$
CNLCU-S †	$6.30_{\pm 2.94}$	$9.62_{\pm 4.67}$	$0.47_{\pm 0.19}$	$7.91_{\pm 2.99}$	$11.76_{\pm 4.61}$	$0.16_{\pm 0.18}$	$8.44_{\pm 2.74}$	$12.66_{\pm 4.30}$	$0.01_{\pm 0.04}$
CNLCU-H †	$6.15_{\pm 2.71}$	$9.46_{\pm 4.47}$	$0.49_{\pm 0.16}$	$7.82_{\pm 3.04}$	$11.85_{\pm 4.59}$	$0.14_{\pm 0.16}$	$8.47_{\pm 2.73}$	$12.67_{\pm 4.23}$	$0.01_{\pm 0.02}$
CoSelfie †	$6.58_{\pm 3.35}$	$9.98_{\pm 5.14}$	$0.43_{\pm 0.24}$	$8.06_{\pm 3.03}$	$12.22_{\pm 4.63}$	$0.10_{\pm 0.13}$	$8.48_{\pm 2.78}$	$12.75_{\pm 4.30}$	$0.00_{\pm 0.01}$
Vanilla (BIN)	$5.23_{\pm 1.87}$	$8.26_{\pm 3.42}$	$0.60_{\pm 0.07}$	$7.77_{\pm 2.87}$	$11.65_{\pm 4.36}$	$0.17_{\pm 0.14}$	$8.40_{\pm 2.82}$	$12.62_{\pm 4.33}$	$0.02_{\pm 0.04}$
SuperLoss (BIN)	$7.30_{\pm 2.47}$	$10.58_{\pm 3.91}$	$0.31_{\pm 0.11}$	$8.26_{\pm 3.01}$	$12.34_{\pm 4.63}$	$0.09_{\pm 0.12}$	$8.37_{\pm 2.82}$	$12.58_{\pm 4.35}$	$0.03_{\pm 0.03}$
RDI (BIN)	$6.11_{\pm 2.18}$	$9.04_{\pm 3.66}$	$0.51_{\pm 0.10}$	$8.24_{\pm 2.65}$	$12.35_{\pm 4.01}$	$0.05_{\pm 0.04}$	$8.35_{\pm 2.79}$	$12.64_{\pm 4.34}$	$0.02_{\pm 0.02}$
AUX (BIN)	$5.72_{\pm 2.30}$	$8.77_{\pm 3.90}$	$0.56_{\pm 0.12}$	$8.15_{\pm 2.44}$	$12.17_{\pm 3.74}$	$0.06_{\pm 0.10}$	$8.25_{\pm 2.87}$	$12.48_{\pm 4.42}$	$0.05_{\pm 0.05}$
C-Mixup (BIN)	$5.87_{\pm 1.95}$	$8.80_{\pm 3.34}$	$0.53_{\pm 0.07}$	$7.88_{\pm 2.79}$	$11.50_{\pm 4.52}$	$0.21_{\pm 0.11}$	$8.06_{\pm 2.70}$	$11.93_{\pm 4.07}$	$0.12_{\pm 0.03}$
CDR (BIN)	$5.92_{\pm 2.20}$	$8.86_{\pm 3.58}$	$0.54_{\pm 0.09}$	$8.04_{\pm 2.71}$	$12.05_{\pm 4.30}$	$0.11_{\pm 0.10}$	$8.41_{\pm 2.82}$	$12.64_{\pm 4.29}$	$0.02_{\pm 0.02}$
Selfie (BIN)	$8.38_{\pm 2.84}$	$12.69_{\pm 4.44}$	$0.02_{\pm 0.03}$	$8.39_{\pm 2.86}$	$12.67_{\pm 4.40}$	$0.02_{\pm 0.03}$	$8.33_{\pm 2.84}$	$12.59_{\pm 4.40}$	$0.03_{\pm 0.05}$
CoTeaching (BIN) †	$8.20_{\pm 2.89}$	$12.33_{\pm 4.35}$	$0.07_{\pm 0.05}$	$8.34_{\pm 2.75}$	$12.52_{\pm 4.23}$	$0.04_{\pm 0.04}$	$8.40_{\pm 2.83}$	$12.67_{\pm 4.35}$	$0.02_{\pm 0.03}$
CNLCU-S (BIN) †	$8.12_{\pm 2.99}$	$12.22_{\pm 4.57}$	$0.10_{\pm 0.12}$	$8.45_{\pm 2.80}$	$12.65_{\pm 4.33}$	$0.02_{\pm 0.03}$	$8.40_{\pm 2.83}$	$12.64_{\pm 4.39}$	$0.02_{\pm 0.03}$
CNLCU-H (BIN) †	$8.22_{\pm 2.72}$	$12.34_{\pm 4.13}$	$0.06_{\pm 0.06}$	$8.35_{\pm 2.83}$	$12.65_{\pm 4.40}$	$0.02_{\pm 0.03}$	$8.37_{\pm 2.82}$	$12.55_{\pm 4.36}$	$0.04_{\pm 0.04}$
CoSelfie (BIN) †	$8.35_{\pm 2.89}$	$12.62_{\pm 4.45}$	$0.03_{\pm 0.04}$	$8.39_{\pm 2.90}$	$12.65_{\pm 4.43}$	$0.02_{\pm 0.04}$	$8.28_{\pm 2.77}$	$12.42_{\pm 4.26}$	$0.05_{\pm 0.06}$
GRAIN (ours)	$\mathbf{4.80_{\pm 1.53}}$	$\mathbf{7.37_{\pm 2.79}}$	$\mathbf{0.64_{\pm 0.06}}$	$\mathbf{6.37_{\pm 1.98}}$	$\mathbf{9.20_{\pm 2.80}}$	$\mathbf{0.45_{\pm 0.21}}$	$\mathbf{7.25_{\pm 2.66}}$	$\mathbf{10.68_{\pm 3.67}}$	$\mathbf{0.27_{\pm 0.12}}$
GRAIN-AUG (ours)	$\mathbf{5.13_{\pm 1.48}}$	$\mathbf{7.70_{\pm 2.84}}$	$\mathbf{0.63_{\pm 0.08}}$	$\mathbf{6.26_{\pm 2.04}}$	$\mathbf{9.15_{\pm 3.19}}$	$\mathbf{0.45_{\pm 0.07}}$	$\mathbf{7.30_{\pm 2.30}}$	$\mathbf{10.94_{\pm 3.19}}$	$\mathbf{0.25_{\pm 0.17}}$

C Discussion: Limitations, Future Work, and Broader Impact

C.1 Connections to Broader Research and Potential Applications

From a statistical perspective, GRAIN relates to measurement-error analysis and likelihood-based inference [1, 34, 35], with broader discussions of noisy-data learning and multivariate regression in [36, 37]. Related work addresses response misclassification [11, 38], covariate errors and variable selection [39, 40, 41, 42], and longitudinal or survival observations with missingness, truncation, or censoring [43, 44, 45, 46]. Extensions to crowdsourced or causation labels [9, 10, 47] and missing or interval-censored responses [48, 49] would require observation-specific models rather than treating all imperfect supervision as additive label noise.

Beyond a single training distribution, source-free adaptation connects uncertainty control, noisy pseudo-labels, and discretized regression [50, 51, 52, 16]. Combining these ideas with structure-preserving graph learning and graph adaptation [53, 54] could extend GRAIN to continuous targets on relational inputs under distribution shift. Polymer-property transfer learning [55] provides a concrete application for investigating such extensions when property measurements are uncertain.

Potential downstream uses include robust auxiliary outcome or score prediction within physician evaluation and treatment-effect analysis [56, 57], longitudinal causal modeling, treatment-rule learning, and behavioral analyses [58, 59]. These applications require task-specific models, with causal identification and confounding addressed separately where appropriate. Evaluation could additionally examine limitations under intentional input perturbations [60, 61, 62].

C.2 Limitations and Outlook

Theoretical Limitations The statistical guarantees in Theorems 4.2 and A.15 are established under a linear feature parameterization over fixed representations, which serves as a tractable proxy for the regression head's behavior. In practice, the feature extractor is a deep network jointly optimized with all other components. One principled connection between these two settings is provided by the lazy training regime [63, 64], under which overparameterized networks trained with small learning rates remain close to their initialization and behave approximately as linear models in the induced reproducing kernel Hilbert space. Under such a regime, the guarantees of Theorems 4.2 and A.15 offer a principled characterization of the regression head's statistical behavior given sufficiently expressive and stable features. We note that the empirical results consistently validate the framework across multiple architectures and noise conditions, providing practical support beyond the linear setting. Bridging this connection rigorously for the joint optimization setting remains an open problem. Additionally, the contrastive regularization in Section 4.2 currently lacks formal guarantees; a rigorous analysis of its order-preserving effect on feature geometry under general noise distributions and jointly optimized representations constitutes an important direction for future theoretical development.

Future Work. We identify several directions for future investigation. First, extending the theoretical guarantees to weakly dependent settings, such as time series or spatially correlated observations, and to more general label noise generation models would broaden the statistical applicability of GRAIN beyond the current i.i.d. assumption; one natural approach is to replace the concentration inequalities used in the current analysis with martingale-based counterparts, such as Azuma–Hoeffding or Bernstein-type inequalities for dependent sequences. Second, rigorously bridging the gap between the linear theoretical setting and the deep network setting used in practice, for instance by leveraging lazy training arguments beyond the regression head to cover the joint optimization of all components, constitutes an important open problem. Third, developing a formal analysis of the contrastive regularization and its order-preserving effect on feature geometry under general noise distributions and jointly optimized representations remains a key theoretical direction; drawing on emerging theory for contrastive learning [65, 66] to characterize how the order-preserving objective shapes the geometry of learned representations under noisy supervision provides a promising starting point.

Broader Impact This work addresses a fundamental challenge in supervised learning: reliable regression under imperfect supervision. The methods developed here are general-purpose and applicable across domains where continuous labels are subject to measurement error or annotation uncertainty, including medical imaging, remote sensing, and scientific instrumentation. In such high-stakes settings, improved robustness to label noise may contribute to more reliable predictive systems and reduce the risk of performance degradation due to data quality issues.

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 claims in the abstract and introduction (Section 1; Figure 2) accurately reflect the paper's contributions and scope, and are supported by both theoretical analysis and empirical results.

   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: The paper discusses the limitations of the proposed method, along with future work directions and broader impact, in Appendix C.

   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: All theoretical results are stated with their full set of assumptions, and complete proofs are provided in Appendix A

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: The paper provides full implementation details in Appendix B.

   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: Dataset information is provided in Appendix B.1, and a pseudo-code description of the training procedure is presented in Algorithm 1. Part of the implementation code is included in the supplemental material. The full codebase will be made publicly available upon acceptance.

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: All training and evaluation details, including data splits, hyperparameter settings, optimizer choice, and selection criteria, are provided in the experimental section (Section 5.1) and the appendix (Appendices B.1 and B.2).

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: All experimental results are reported as means with standard errors over multiple independent runs with different random seeds. Tables in Section 5.2 and Appendix B.3 include standard errors.

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: The computational resources used in our experiments, including GPU type, memory, and approximate training time, are reported in Appendix B.

   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: Yes. This research conforms with the NeurIPS Code of Ethics in every respect. The work involves only publicly available benchmark datasets, no human subjects or private data were used, and we foresee no harmful applications of the proposed method.

   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: [Yes]

    Justification: A detailed discussion is provided in Appendix C.

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: This paper does not release any pre-trained models, image generators, or scraped datasets. The work proposes a training methodology evaluated on publicly available benchmark datasets, and thus no special safeguards for data or model release are required.

    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 datasets and code assets used in this paper are publicly available benchmarks that are properly cited in the references in Appendix B.1.

    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: [Yes]

    Justification: The code implementing the proposed GRAIN framework will be made publicly available upon acceptance, with documentation covering installation, dataset preparation, and reproduction of the experimental results reported in the paper.

    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: This paper does not involve any crowdsourcing experiments or 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: This paper does not involve any research with human subjects, and therefore no IRB approval or equivalent ethical review was required.

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: No LLMs were used as a component of the core methodology. LLMs were used only for writing and editing assistance, which does not require declaration per the NeurIPS guidelines.

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.