CoRe: Coarse-to-Fine Learning for Few-Shot Multimodal Sentiment Regression

Anonymous authors
Paper under double-blind review

Abstract

Multimodal sentiment regression (MSR) integrates complementary information across modalities to model sentiment polarity and intensity in a continuous sentiment space. However, existing MSR methods typically rely on large-scale annotated datasets, which are costly and difficult to obtain. While few-shot MSR offers a promising solution, it still faces three key challenges: (i) limited training sets may fail to preserve the underlying continuous sentiment distribution; (ii) pretrained models are typically better at semantic understanding than at numerical regression; and (iii) conventional regression losses focus on point-wise errors, overlooking distributional discrepancies. To address these challenges, we make the first attempt to investigate few-shot MSR and propose CoRe, a coarse-to-fine learning framework. Specifically, CoRe introduces a continuous distribution sampling strategy to construct a representative few-shot training set that closely approximates the global sentiment distribution. It further reformulates sentiment regression as a coarse-to-fine prediction process, progressively refining pretrained affective semantics into continuous sentiment predictions, with a heterogeneous multimodal graph capturing cross-modal and contextual dependencies. Moreover, CoRe employs a distance-aware objective to align predicted sentiment distributions with soft targets, thereby reducing distributional discrepancies. Extensive experiments on four datasets demonstrate that CoRe outperforms state-of-the-art baselines, reducing MAE by 6.13% and improving Corr by 5.03% on average.

1 Introduction

Multimodal sentiment analysis is an important research topic in affective computing, aiming to infer users' sentiment from multimodal signals such as text, video, and audio (Zheng et al., 2026). Many existing methods formulate sentiment analysis as a classification task; however, even fine-grained classification discretizes the continuous sentiment space into predefined categories, making it difficult to quantify differences in sentiment (Jing et al., 2026). In contrast, Multimodal Sentiment Regression (MSR) models sentiment in a continuous space, explicitly representing sentiment polarity and intensity while enabling accurate quantification of these differences (Nicolaou et al., 2011; Paltoglou & Thelwall, 2013). This continuous and measurable modeling is particularly valuable for applications requiring quantitative estimation of affective states, including mental health assessment, depression monitoring, and human-computer interaction (Qiu et al., 2025; Tellamekala et al., 2024).

However, existing MSR methods typically rely on large-scale annotated datasets, which are costly and difficult to obtain (Liu et al., 2026). Few-Shot Learning (FSL) provides a promising way to reduce this reliance and has recently been applied to multimodal sentiment analysis (Huang & Michailidis, 2026). Existing studies such as PVLM (Yu & Zhang, 2022), UP-MPF (Yu et al., 2022), and MultiPoint (Yang et al., 2023b) mainly address few-shot sentiment classification in image-text settings, while GMP (Yang et al., 2023a) and VERO (Sun et al., 2025) investigate few-shot multimodal aspect-based sentiment classification. Despite this progress, existing few-shot studies in this field remain primarily focused on discrete classification, leaving few-shot MSR largely unexplored.

A key challenge in few-shot MSR is constructing a representative few-shot training set that approximates the global sentiment distribution of the full training set, as illustrated in Figure 1(a). In classification, predefined categories partition the label space, allowing samples to be selected within each class and the class distribution to be controlled (Li et al., 2024; Jia et al., 2025). In regres-

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

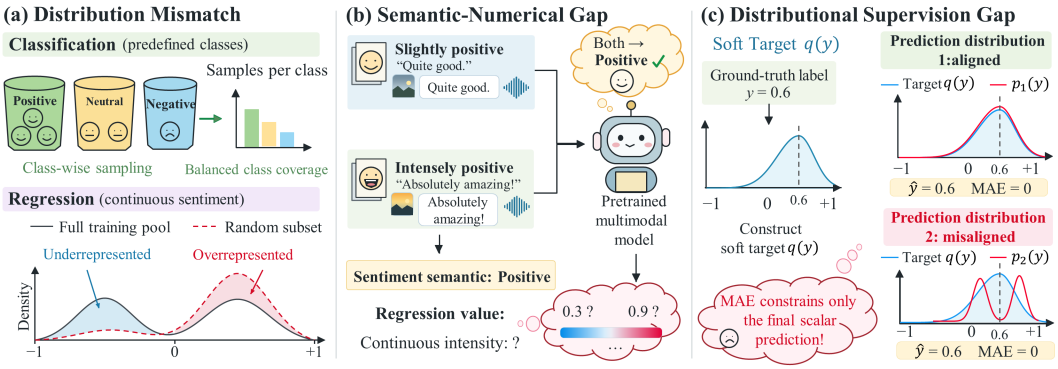


Figure 1: Challenges in few-shot multimodal sentiment regression.

sion, however, labels lie in a continuous space without predefined category boundaries, making it difficult to control the distribution of a small training subset (Guo et al., 2025). Although random sampling can produce a few-shot training set, its label distribution may deviate substantially from that of the full training set (Nie et al., 2025). This motivates a sampling strategy that preserves the global sentiment distribution under a limited sample budget.

Meanwhile, existing FSL methods often rely on pretrained models, which provide strong semantic and cross-modal representations but are generally better at semantic understanding than at numerical regression (Jung et al., 2026). As illustrated in Figure 1(b), pretrained models can readily recognize sentiment polarity, whereas MSR further requires distinguishing subtle differences in sentiment intensity and producing precise continuous sentiment estimates. Under limited supervision, adapting pretrained models from semantic understanding to continuous regression remains difficult (Du et al., 2024; Li et al., 2022). This motivates a coarse-to-fine mechanism that progressively transforms pretrained semantic representations into continuous sentiment predictions.

In addition, conventional regression losses such as MAE focus on point-wise errors, leaving the predicted distribution unconstrained (Zhang et al., 2023b; Li et al., 2022). As illustrated in Figure 1(c), different predicted distributions can yield the same scalar prediction and thus the same MAE, even though one closely matches the target distribution while the other does not. Such point-wise supervision therefore overlooks distributional discrepancies between the predicted and target distributions (Cardoso et al., 2025). This motivates an additional distribution-level objective that explicitly aligns the predicted distribution with the target distribution.

To address these challenges, we make the first attempt to investigate few-shot MSR and propose CoRe, a coarse-to-fine learning framework. Specifically, CoRe estimates the global sentiment distribution from zero-shot predictions of a pretrained multimodal model and constructs a representative few-shot training set by selecting a limited subset whose distribution closely approximates the estimated global distribution. CoRe further reformulates sentiment regression as a coarse-to-fine prediction process that leverages the affective semantics encoded in pretrained models and progressively refines them into continuous sentiment predictions, with a heterogeneous multimodal graph capturing cross-modal and contextual dependencies. Finally, CoRe introduces a distance-aware objective that complements point-wise regression supervision by aligning the predicted distribution with the corresponding soft target distribution, thereby reducing distributional discrepancies.

We make the following main contributions: 1) We investigate few-shot MSR and propose CoRe, a coarse-to-fine learning framework. To the best of our knowledge, this is the first attempt to study this setting. 2) We propose a continuous distribution sampling strategy that constructs a representative few-shot training set by approximating the global sentiment distribution. 3) We develop a coarse-to-fine learning strategy that transforms pretrained affective semantics into continuous sentiment predictions, together with a distance-aware objective for distribution-level supervision. 4) Extensive experiments on four datasets show that CoRe consistently outperforms state-of-the-art baselines.

2 Related Work

Multimodal Sentiment Regression. MSR is a core task in multimodal sentiment analysis, aiming to infer continuous sentiment values from language, vision, and audio. Compared with discrete

sentiment classification, MSR models sentiment polarity and intensity on a continuous scale (Xu et al., 2026; Luo et al., 2025; Wang et al., 2025). Existing MSR methods have explored tensor-based fusion (Zadeh et al., 2017; Liu et al., 2018), cross-modal attention (Tsai et al., 2019a; Yu et al., 2021), and modality-invariant representation learning (Hazarika et al., 2020; Wang et al., 2023b). Recent studies further exploit label relationships and pretrained multimodal representations to improve sentiment regression (Yang et al., 2025; Zhuang et al., 2024; Huang et al., 2026; Zhuang et al., 2026b). However, these methods are predominantly developed under fully supervised settings with sufficient annotations, limiting their applicability when labeled data are scarce.

Few-Shot Multimodal Sentiment Analysis. Recent studies explore FSL for multimodal sentiment analysis to reduce annotation dependence, but primarily focus on discrete prediction. PVLM (Yu & Zhang, 2022), UP-MPF (Yu et al., 2022), and MultiPoint (Yang et al., 2023b) investigate few-shot sentiment classification in image-text settings, while GMP (Yang et al., 2023a) and VERO (Sun et al., 2025) extend FSL to multimodal aspect-based sentiment classification. Beyond sentiment classification, few-shot regression has been explored in other domains through meta-learning and probabilistic modeling (Sendera et al., 2021; Iwata et al., 2025; Shi et al., 2024; Zhao & Chen, 2020), while vision-language methods investigate ordinal regression over ordered discrete targets (Du et al., 2024; Wang et al., 2023c). However, these existing approaches are designed for different task settings and are not readily applicable to MSR, leaving few-shot MSR largely unexplored.

3 METHOD

We propose CoRe, a coarse-to-fine learning framework for few-shot MSR. As illustrated in Figure 2, CoRe consists of four components: (1) continuous distribution sampling for constructing a representative few-shot training set; (2) heterogeneous multimodal graph modeling for capturing cross-modal and contextual dependencies; (3) coarse-to-fine sentiment prediction for refining pretrained affective semantics into continuous sentiment predictions; and (4) distance-aware supervision for reducing discrepancies between the predicted and target sentiment distributions.

3.1 PROBLEM FORMULATION

Given a multimodal dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, each sample is represented as $X_i = (x_i^t, x_i^a, x_i^v)$, where x_i^t, x_i^a, and x_i^v denote the textual, audio, and visual modalities. $y_i \in \mathcal{Y} \subset \mathbb{R}$ is the continuous sentiment label. MSR aims to learn a mapping $f_\theta : \mathcal{X}^t \times \mathcal{X}^a \times \mathcal{X}^v \to \mathcal{Y}$ such that $\hat{y}_i = f_\theta(x_i^t, x_i^a, x_i^v)$ approximates y_i. Let $\mathcal{D}_{\text{train}} \subset \mathcal{D}$ denote the training set. In the few-shot setting, only a limited subset $\mathcal{D}_{\text{few}} \subset \mathcal{D}_{\text{train}}$ with $|\mathcal{D}_{\text{few}}| = K$ is used for training. Our goal is to learn f_θ from $\mathcal{D}_{\text{few}}$ and produce accurate continuous sentiment predictions for unseen multimodal samples.

3.2 CONTINUOUS DISTRIBUTION SAMPLING

Given the training set $\mathcal{D}_{\text{train}} = \{X_i\}_{i=1}^{N_t}$, we first use a pretrained multimodal model to obtain a zero-shot sentiment prediction $\tilde{y}_i$ for each training sample. These predictions serve as a proxy for the underlying global sentiment distribution. We estimate the global sentiment density via Kernel Density Estimation (KDE):

$$p(y) = \frac{1}{N_t h} \sum_{i=1}^{N_t} \kappa\left(\frac{y - \tilde{y}_i}{h}\right) \tag{1}$$

where $\kappa(\cdot)$ denotes the kernel function and h is the bandwidth.

We then partition the continuous sentiment space into B intervals $\{\mathcal{I}_b\}_{b=1}^{B}$ and compute the probability mass of each interval:

$$p_b = \int_{\mathcal{I}_b} p(y)\, dy, \qquad \boldsymbol{p} = [p_1, \ldots, p_B] \tag{2}$$

where p_b denotes the probability mass of interval $\mathcal{I}_b$, and $\boldsymbol{p}$ represents the estimated global sentiment distribution. For a candidate subset $\mathcal{S}$, we denote its distribution over the same intervals by $\boldsymbol{q}_{\mathcal{S}}$. The objective of few-shot subset selection can then be formulated as:

$$\mathcal{S}^* = \arg\min_{\mathcal{S} \in \Omega_K} D\left(\boldsymbol{q}_{\mathcal{S}}, \boldsymbol{p}\right) \tag{3}$$

where Ω_K denotes the set of feasible subsets under the sampling budget K, and $D(\cdot, \cdot)$ measures the discrepancy between the candidate and target distributions.

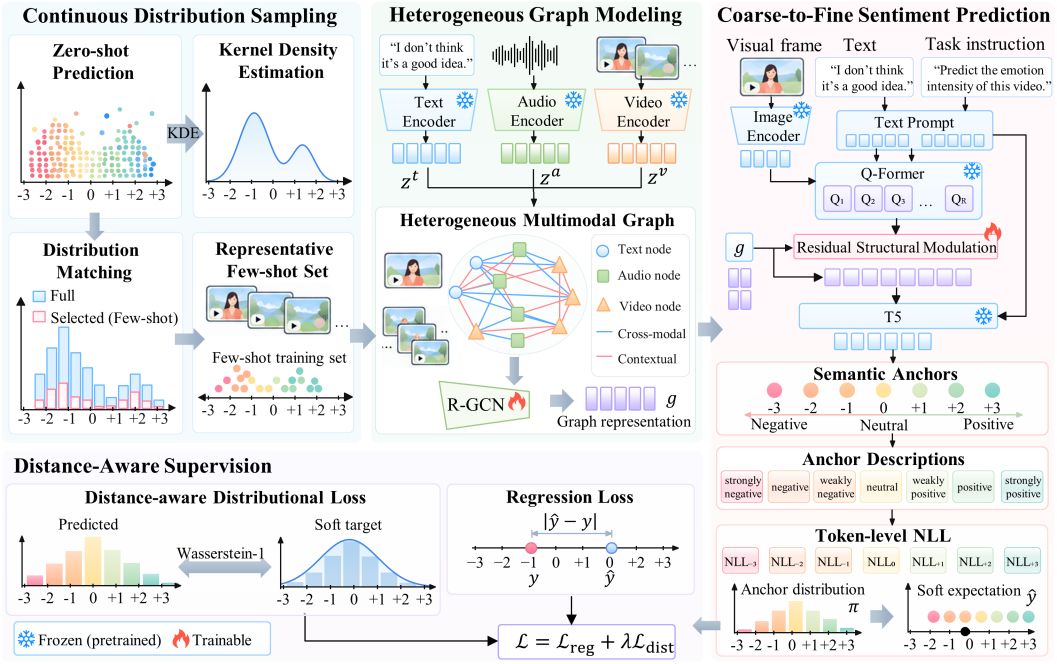


Figure 2: Overview of the proposed CoRe framework for few-shot MSR.

The selection strategy depends on the sampling unit. For segment-level sampling, $q_{\mathcal{S},b} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \mathbf{1}[\tilde{y}_i \in \mathcal{I}_b]$, and the sampling budget assigned to interval $\mathcal{I}_b$ is determined by $k_b \approx K p_b$. For video-level sampling, all segments from a selected video are retained together, and $\boldsymbol{q}_{\mathcal{S}}$ is computed from the zero-shot predictions of all segments contained in the selected videos. We measure its discrepancy from the target distribution using the Wasserstein-1 distance:

$$W_1(\boldsymbol{p}, \boldsymbol{q}) = \sum_{b=1}^{B-1} |F_{\boldsymbol{p}}(b) - F_{\boldsymbol{q}}(b)|\, \Delta_b \tag{4}$$

where $F_{\boldsymbol{p}}(b) = \sum_{r=1}^{b} p_r$ and $F_{\boldsymbol{q}}(b) = \sum_{r=1}^{b} q_r$ denote the cumulative probability masses, and Δ_b is the distance between the centers of adjacent intervals. Further details are in Appendix A.10.

3.3 Heterogeneous Multimodal Graph Modeling

Given a target sample X_i, let $\boldsymbol{z}_i^m$ denote its projected feature for modality $m \in \mathcal{M}$. To capture both cross-modal interactions within each sample and contextual dependencies between the target sample and its relevant neighbors, we construct a context-aware heterogeneous multimodal graph.

Specifically, we first retrieve relevant contextual samples from candidates belonging to the same video. For each candidate sample j, its multimodal similarity to the target sample is computed by averaging the non-negative cosine similarities across modalities:

$$s_{ij} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \max\left(0, \frac{(\boldsymbol{z}_i^m)^\top \boldsymbol{z}_j^m}{\|\boldsymbol{z}_i^m\|_2 \|\boldsymbol{z}_j^m\|_2}\right) \tag{5}$$

The top-k candidates according to s_{ij} are retained to form the contextual neighborhood $\mathcal{N}_i$.

Feature similarity alone may not fully characterize the relevance of a contextual relation. We therefore introduce a learnable relation gate. The modality features are concatenated as $\boldsymbol{z}_i = [\boldsymbol{z}_i^t; \boldsymbol{z}_i^a; \boldsymbol{z}_i^v]$ and further projected into a relation space, yielding $\bar{\boldsymbol{z}}_i$. For each target–context pair (i, j), the relation gate is computed as:

$$g_{ij} = \sigma\left(f_{\text{rel}}\left([\bar{\boldsymbol{z}}_i; \bar{\boldsymbol{z}}_j; \bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}}_j; \bar{\boldsymbol{z}}_i \odot \bar{\boldsymbol{z}}_j]\right)\right) \tag{6}$$

where $f_{\text{rel}}(\cdot)$ is a learnable MLP. The final contextual relation weight is defined as $\omega_{ij} = s_{ij} g_{ij}$, combining the similarity prior with the learned relation gate.

The target sample and its contextual neighbors are then expanded into modality-specific text, audio, and visual nodes. We consider two relation types: cross-modal relations connect different modalities within the same sample, while contextual relations connect the target sample with its contextual

neighbors. The contextual relation weights ω_{ij} are assigned to the corresponding contextual edges and incorporated into the relation-specific adjacency matrices. Let $\boldsymbol{H}^{(l)}$ denote the node representations at the l-th graph layer, $\widetilde{\boldsymbol{A}}_r$ the normalized adjacency matrix for relation $r \in \mathcal{R}$, and $\boldsymbol{W}_r^{(l)}$ the corresponding relation-specific transformation. Relation-aware graph propagation is defined as:

$$\boldsymbol{Z}^{(l)} = \sum_{r\in\mathcal{R}} \widetilde{\boldsymbol{A}}_r \boldsymbol{H}^{(l)} \boldsymbol{W}_r^{(l)} + \boldsymbol{b}^{(l)} \tag{7}$$

The node representations are then updated through Eq. 8:

$$\boldsymbol{H}^{(l+1)} = \boldsymbol{H}^{(l)} + \mathrm{Dropout}\left(\zeta\left(\mathrm{LN}\left(\boldsymbol{Z}^{(l)}\right)\right)\right) \tag{8}$$

where $\zeta(\cdot)$ denotes the nonlinear activation function. After L graph layers, the target-specific text, audio, and visual node representations are aggregated as $\boldsymbol{g}_i = \mathrm{Readout}(\boldsymbol{h}_{i,t}^{(L)}, \boldsymbol{h}_{i,a}^{(L)}, \boldsymbol{h}_{i,v}^{(L)})$. The resulting structural representation $\boldsymbol{g}_i$ is subsequently used to modulate the pretrained query representations for coarse-to-fine sentiment prediction. Further details are provided in Appendix A.11.

3.4 Coarse-to-Fine Sentiment Prediction

To bridge pretrained sentiment semantics and continuous sentiment values under limited supervision, we reformulate MSR as a coarse-to-fine semantic prediction process. Specifically, for each target sample X_i, its visual input and textual prompt are processed by InstructBLIP to obtain the Q-Former query representations $\boldsymbol{U}_i = [\boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,R}]$. The target-specific structural representation $\boldsymbol{g}_i$ from the heterogeneous graph is then injected into these queries through residual modulation:

$$\widetilde{\boldsymbol{u}}_{i,r} = \boldsymbol{u}_{i,r} + \rho\, \psi(\boldsymbol{g}_i), \qquad \rho = \sigma(\alpha), \qquad r = 1, \ldots, R \tag{9}$$

where $\psi(\cdot)$ projects $\boldsymbol{g}_i$ into the query representation space, $\sigma(\cdot)$ is the sigmoid function, and α is a learnable scalar controlling the modulation strength. The modulated queries, together with the textual prompt, are further processed by T5 to obtain the sample representation $\boldsymbol{e}_i$.

We then introduce C semantic anchors to establish the correspondence between sentiment semantics and regression values. The anchors are uniformly defined over the continuous sentiment space:

$$\mathcal{A} = \{a_c\}_{c=1}^{C}, \qquad a_c = y_{\min} + \frac{c-1}{C-1}\,(y_{\max} - y_{\min}) \tag{10}$$

where each anchor a_c is associated with a predefined textual description s_c characterizing the corresponding sentiment polarity and intensity.

Given the c-th semantic description $s_c = (w_{c,1}, \ldots, w_{c,L_c})$, where L_c denotes the number of valid output tokens, we measure its compatibility with the current sample using the length-normalized token-level negative log-likelihood (NLL):

$$\ell_{i,c} = -\frac{1}{L_c} \sum_{l=1}^{L_c} \log P_\phi\left(w_{c,l} \mid w_{c,<l}, \boldsymbol{e}_i\right) \tag{11}$$

where ϕ denotes the model parameters. A smaller $\ell_{i,c}$ indicates greater compatibility with the corresponding sentiment semantics. The negative NLL values are normalized across all anchors:

$$\pi_{i,c} = \frac{\exp(-\ell_{i,c})}{\sum_{r=1}^{C} \exp(-\ell_{i,r})} \tag{12}$$

yielding the semantic distribution $\boldsymbol{\pi}_i = [\pi_{i,1}, \ldots, \pi_{i,C}]^\top$. The semantic anchors provide coarse sentiment semantics, while soft expectation converts the resulting semantic distribution into a continuous prediction. Rather than selecting a single anchor, we retain the full distribution and compute:

$$\hat{y}_i = \sum_{c=1}^{C} \pi_{i,c} a_c \tag{13}$$

Since $\hat{y}_i$ is a convex combination of the anchors, it remains within the valid sentiment range without being restricted to discrete anchor values. More analysis is provided in Appendix A.12.

3.5 Distance-Aware Supervision

The coarse-to-fine prediction produces both a continuous prediction $\hat{y}_i$ and a semantic distribution $\boldsymbol{\pi}_i$ over the ordered anchors. Supervising only the scalar prediction does not fully constrain the predicted distribution, since different anchor distributions may yield identical or similar expectations. To incorporate the ordering and distance information among semantic anchors into supervision, we construct a soft target distribution over the same anchor set $\mathcal{A} = \{a_c\}_{c=1}^{C}$ from the ground-truth value y_i. Specifically, the target probability assigned to anchor a_c is defined as:

$$\tau_{i,c} = \frac{\exp\left(-\frac{(a_c - y_i)^2}{2\gamma^2}\right)}{\sum_{r=1}^{C} \exp\left(-\frac{(a_r - y_i)^2}{2\gamma^2}\right)} \tag{14}$$

Here, γ controls the smoothness of the target distribution and is set to the spacing between adjacent anchors. The resulting soft target is denoted by $\boldsymbol{\tau}_i = [\tau_{i,1}, \ldots, \tau_{i,C}]^\top$. Unlike hard assignment to a single anchor, the soft target assigns larger probabilities to anchors closer to y_i, thereby encoding the numerical proximity between the ground-truth value and the semantic anchors.

We then measure the discrepancy between the predicted distribution $\boldsymbol{\pi}_i$ and the target distribution $\boldsymbol{\tau}_i$ using the Wasserstein-1 distance. Let $F_{\boldsymbol{\pi}_i}(c) = \sum_{r=1}^{c} \pi_{i,r}$ and $F_{\boldsymbol{\tau}_i}(c) = \sum_{r=1}^{c} \tau_{i,r}$ denote their cumulative probability masses. The distance-aware distributional loss is defined as:

$$\mathcal{L}_{\text{dist}}^{(i)} = \sum_{c=1}^{C-1} \left|F_{\boldsymbol{\pi}_i}(c) - F_{\boldsymbol{\tau}_i}(c)\right| (a_{c+1} - a_c) \tag{15}$$

This objective accounts for both probability-mass mismatch and the distance over which the mass is displaced, assigning larger penalties to distributional errors spanning greater distances. We further retain direct supervision on the continuous prediction:

$$\mathcal{L}_{\text{reg}}^{(i)} = |\hat{y}_i - y_i| \tag{16}$$

The final training objective combines scalar regression and distance-aware supervision:

$$\mathcal{L}_i = \mathcal{L}_{\text{reg}}^{(i)} + \lambda \mathcal{L}_{\text{dist}}^{(i)} \tag{17}$$

where λ controls the contribution of the distributional loss. The objective is averaged over all samples in each mini-batch. The $\mathcal{L}_{\text{reg}}$ term directly supervises the final continuous prediction, while the $\mathcal{L}_{\text{dist}}$ term penalizes distributional discrepancies according to the distances between semantic anchors. Further theoretical analysis is provided in Appendix A.13.

Complexity analysis. We analyze the additional computational cost of CoRe beyond the frozen backbone. Let M be the number of context candidates, D the multimodal feature dimension, h the graph hidden dimension, and d the final representation dimension. The additional time complexity is $O\big(K(MD + Dh + h^2 + hd)\big)$. Detailed efficiency analysis is provided in Appendix A.8.

4 Experiments

Datasets. We evaluate CoRe on four widely used multimodal sentiment benchmarks: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018b), CH-SIMS (Yu et al., 2020), and CH-SIMS v2 (Liu et al., 2022). CMU-MOSI and CMU-MOSEI use sentiment labels in $[-3, 3]$, while CH-SIMS and CH-SIMS v2 use sentiment labels in $[-1, 1]$. All experiments use a 7% few-shot split of the full training set. For CMU-MOSI and CMU-MOSEI, videos are treated as the sampling units, with all segments from each selected video retained. For CH-SIMS and CH-SIMS v2, segments are sampled individually. Detailed dataset statistics and split information are provided in Appendix A.1.

Baselines. We compare CoRe with three groups of baselines. Multimodal regression methods include TFN (Zadeh et al., 2017), TETFN (Wang et al., 2023a), EF-LSTM (Williams et al., 2018), MoMKE (Xu et al., 2024), IIE (Zhuang et al., 2026b), KAN-MCP (Luo et al., 2025), TMDC (Zhuang et al., 2026a), MMIM (Han et al., 2021), MFN (Zadeh et al., 2018a), MISA (Hazarika et al., 2020), LMF (Liu et al., 2018), Self-MM (Yu et al., 2021), LF-DNN (Williams et al., 2018), MFM (Tsai et al., 2019b), Graph-MFN (Zadeh et al., 2018b), CENET (Wang et al., 2023b), ALMT (Zhang et al., 2023a), GLoMo (Zhuang et al., 2024), DLF (Wang et al., 2025), MulT (Tsai et al., 2019a), RECAP (Huang et al., 2026), and MSE-Adapter (Yang et al., 2025). Few-shot multimodal methods include PVLM (Yu & Zhang, 2022) and VERO (Sun et al., 2025), which are adapted to continuous regression with regression heads. Pretrained-model baselines include zero-shot InstructBLIP with the proposed coarse-to-fine sentiment prediction, and InstructBLIP equipped with a regression head. All trainable baselines use the same few-shot training subsets and the same validation and test splits.

Evaluation Metrics. Following prior work (Yu et al., 2021), we report MAE and Corr as regression metrics and Acc/F1 as discretized classification metrics, using the standard discretization protocols of the corresponding benchmarks. For CMU-MOSI and CMU-MOSEI, both Acc_2 and F1 are reported under negative/non-negative and negative/positive settings. We additionally report Acc_5 and Acc_7 for CMU-MOSI and CMU-MOSEI, and Acc_3 and Acc_5 for CH-SIMS and CH-SIMS v2.

Experimental Settings. All experiments are conducted on an Ubuntu system equipped with an NVIDIA RTX A6000 GPU. We use InstructBLIP (Dai et al., 2023) with Flan-T5-XL as the default backbone and additionally report results with Flan-T5-XXL (Chung et al., 2024) in the main comparison. The pretrained InstructBLIP parameters are kept frozen, and only the newly introduced

Table 1: Comparison results on CMU-MOSI and CMU-MOSEI. Acc and F1 are reported in %.

Method	CMU-MOSI						CMU-MOSEI					
	MAE ↓	Corr ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	F1 ↑	MAE ↓	Corr ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	F1 ↑
TFN	1.145	0.560	68.08/68.81	33.24	29.53	67.92/68.74	0.740	0.490	71.83/74.13	42.78	42.60	72.69/74.13
TETFN	1.434	0.037	48.40/47.16	15.45	15.45	34.75/33.62	0.642	0.650	77.76/80.84	47.53	46.66	78.39/80.83
Self-MM	0.930	0.714	77.09/78.32	39.30	34.37	77.11/78.42	0.609	0.693	78.50/82.07	49.53	48.47	79.20/82.14
MulT	1.313	0.439	59.30/58.87	24.29	22.71	55.28/54.95	0.731	0.526	74.05/76.96	42.09	41.92	74.78/76.89
MMIM	0.932	0.721	77.32/78.41	40.90	35.22	77.34/78.50	0.616	0.684	79.57/82.19	48.98	48.06	80.00/82.08
MISA	0.998	0.678	74.46/75.82	34.81	31.28	74.41/75.87	0.643	0.661	78.44/81.56	46.55	45.71	79.05/81.55
MFN	1.172	0.513	67.46/68.41	30.32	27.99	67.33/68.37	0.712	0.575	72.76/77.14	42.94	42.55	73.88/77.39
MFM	1.162	0.534	67.93/68.75	30.70	28.04	67.94/68.86	0.741	0.482	71.37/74.23	42.16	42.07	72.33/74.29
LMF	1.211	0.405	64.34/65.21	28.10	25.75	64.31/65.28	0.725	0.531	73.56/76.76	43.45	43.37	74.40/76.78
LF-DNN	1.183	0.532	68.34/69.09	30.20	27.67	68.31/69.16	0.719	0.545	75.14/77.49	43.18	42.94	75.76/77.39
Graph-MFN	1.131	0.572	68.69/69.36	32.65	29.15	68.60/69.37	0.721	0.535	75.91/77.43	42.66	42.41	76.20/77.05
EF-LSTM	1.243	0.514	65.66/66.07	24.61	23.79	62.22/62.70	0.753	0.456	72.77/73.78	42.40	42.28	72.98/73.14
CENET	1.070	0.635	72.65/73.47	32.54	29.62	72.63/73.54	0.657	0.655	76.90/81.10	45.70	44.73	77.79/81.27
ALMT	0.973	0.693	76.27/77.59	34.37	30.96	76.30/77.69	0.615	0.690	79.26/82.33	48.11	47.16	79.79/82.28
MSE-Adapter (6B)	0.853	0.728	82.94/83.32	46.88	39.01	82.94/83.27	0.681	0.606	82.94/64.37	45.70	44.36	82.39/64.13
KAN-MCP	1.249	0.481	68.88/69.40	22.89	22.69	68.65/69.28	0.833	0.142	71.29/64.03	41.37	41.37	61.28/52.01
GLoMo	1.233	0.527	72.82/73.99	19.94	19.94	72.80/74.05	0.649	0.638	77.57/80.09	47.19	46.30	78.12/80.02
IIE	0.998	0.693	75.95/77.10	34.34	31.63	75.85/77.09	0.607	0.697	79.14/82.55	49.92	48.81	79.76/82.57
DLF	1.431	0.015	50.82/51.25	15.45	15.45	35.83/36.51	0.647	0.643	77.81/80.56	47.41	46.71	78.41/80.54
MoMKE	1.052	0.639	74.72/75.46	31.98	29.59	74.70/75.52	0.683	0.620	77.95/81.30	43.99	43.78	78.57/81.29
TMDC	0.947	0.693	79.39/81.01	35.69	32.86	79.39/81.08	0.663	0.675	78.50/82.76	45.02	43.86	79.24/82.83
RECAP	1.225	0.391	64.84/65.79	29.80	26.88	64.73/65.48	0.812	0.221	71.68/66.19	41.56	41.50	64.10/58.23
PVLM	0.884	0.743	78.48/79.82	41.37	35.63	78.53/79.93	0.627	0.673	77.11/81.36	47.97	46.95	77.92/81.45
VERO(LLaVA-7B)	0.757	0.801	84.72/86.31	48.77	41.72	84.72/86.35	0.634	0.699	78.88/83.27	46.18	44.91	79.62/83.36
VERO(LLaVA-13B)	0.754	0.784	83.53/85.15	49.59	41.78	83.45/85.13	0.586	0.740	80.11/84.66	49.85	48.48	80.74/84.68
InstructBLIP(xl)	0.790	0.805	83.53/84.15	54.52	44.46	83.57/84.24	0.650	0.706	81.86/86.62	47.66	46.22	82.50/86.71
InstructBLIP(xxl)	0.645	0.848	86.15/87.65	58.02	48.69	86.14/87.68	0.601	0.773	85.19/86.87	49.98	47.08	85.26/86.67
InstructBLIP(reg)	1.146	0.506	69.65/71.01	29.56	27.90	69.71/71.16	0.581	0.747	78.85/85.29	48.54	47.87	79.75/85.46
CoRe(xl)	0.619	0.855	86.79/88.57	55.45	48.25	86.75/88.56	0.560	0.757	81.82/86.18	51.80	50.35	82.38/86.21
CoRe(xxl)	**0.558**	**0.876**	**87.06/89.33**	**62.24**	**52.77**	**87.01/89.33**	**0.537**	**0.780**	**84.00/87.50**	**53.99**	**52.66**	**84.40/87.49**

CoRe modules are optimized. Text, audio, and video features are extracted using DeBERTa-large (He et al., 2021), wav2vec-large (Baevski et al., 2020), and MA-Net (Zhao et al., 2021), respectively. Models are optimized with AdamW using a weight decay of 1×10^{-4}. The learning rate is set to 3×10^{-4} for CMU-MOSI, CMU-MOSEI, and CH-SIMS v2, and 1×10^{-3} for CH-SIMS. The graph encoder consists of two R-GCN (Schlichtkrull et al., 2018) layers with a hidden dimension of 512. Each batch contains all selected segments from one video in the few-shot training subset. We use seven semantic anchors, set $\gamma = 1$ for CMU-MOSI/CMU-MOSEI and $\gamma = 1/3$ for CH-SIMS/CH-SIMS v2, use a context size of 3, estimate the KDE bandwidth h separately for each dataset using Scott's rule, and set $\lambda = 0.5$. The number of intervals B is set to 13, 58, 10, and 10 for CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMS v2, respectively. Training is performed with bfloat16 mixed precision. Each experiment is repeated five times with different random seeds, and the mean performance is reported, with the corresponding standard deviations provided in Appendix A.2. Additional implementation details, including the prompts and semantic anchor configurations, are provided in Appendix A.7.

4.1 Comparative Analysis

Tables 1 and 2 present a comprehensive comparison of CoRe with existing methods on four benchmark datasets. Overall, CoRe achieves the best performance on the two primary regression metrics, MAE and Corr, while remaining highly competitive on the discretized classification metrics. Compared with the strongest baseline on each dataset, CoRe reduces MAE by 6.13% and improves Corr by 5.03% on average across the four datasets. Compared with conventional multimodal regression methods, CoRe consistently achieves better regression performance under the same few-shot setting, demonstrating its effectiveness in learning continuous sentiment representations from limited supervision. CoRe also outperforms few-shot multimodal baselines, indicating that methods designed primarily for discrete sentiment prediction are less effective for continuous sentiment regression. The zero-shot InstructBLIP results show that pretrained multimodal models provide useful affective priors, while adding a trainable regression head alone yields limited and inconsistent improvements. In contrast, CoRe more effectively adapts pretrained affective semantics to continuous sentiment regression through coarse-to-fine prediction, heterogeneous graph modeling, and distance-aware supervision. We further compare CoRe under the 7% few-shot setting with fully supervised baselines trained on the full training set in Appendix A.9. Overall, these results demonstrate the effectiveness of CoRe for few-shot multimodal sentiment regression.

Table 2: Comparison results on CH-SIMS and CH-SIMS v2. Acc and F1 are reported in %.

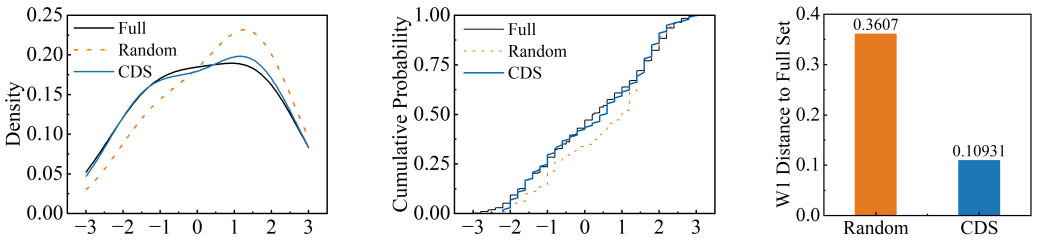
Method	CH-SIMS						CH-SIMS v2					
	MAE ↓	Corr ↑	Acc_2 ↑	Acc_3 ↑	Acc_5 ↑	F1 ↑	MAE ↓	Corr ↑	Acc_2 ↑	Acc_3 ↑	Acc_5 ↑	F1 ↑
TFN	0.535	0.358	68.71	54.18	24.72	68.06	0.443	0.402	68.30	57.95	39.01	67.97
TETFN	0.608	0.015	54.40	24.46	16.54	43.19	0.511	0.036	47.33	15.59	12.88	40.91
Self-MM	0.632	0.021	49.28	28.75	16.24	37.89	0.359	0.604	74.82	67.19	46.42	74.86
MulT	0.701	0.004	48.40	36.76	18.25	43.45	0.375	0.594	74.68	64.86	43.08	74.76
MMIM	0.497	0.476	74.97	59.83	30.94	73.50	0.371	0.596	75.57	67.29	45.28	75.50
MISA	0.507	0.437	71.95	56.37	26.65	71.34	0.423	0.436	68.34	53.54	36.87	68.21
MFN	0.519	0.375	69.32	55.80	28.53	68.64	0.395	0.525	72.56	60.64	41.22	72.27
MFM	0.588	0.099	69.37	54.23	21.19	56.82	0.510	0.011	54.64	16.98	13.83	38.84
LMF	0.545	0.287	70.28	48.10	23.94	68.58	0.437	0.426	69.50	56.31	37.89	69.02
LF-DNN	0.511	0.421	72.69	58.73	30.37	72.20	0.427	0.470	68.94	59.25	40.08	69.05
Graph-MFN	0.516	0.411	71.38	56.46	27.57	70.49	0.435	0.379	67.27	60.35	40.29	64.03
EF-LSTM	0.598	0.023	69.37	30.77	17.55	56.82	0.400	0.526	72.38	63.33	41.99	72.05
CENET	0.598	0.051	64.77	45.91	20.79	59.98	0.375	0.603	75.58	66.44	44.20	75.43
ALMT	0.520	0.385	71.12	55.01	28.10	70.31	0.386	0.549	72.46	62.30	42.09	72.53
MSE-Adapter(6B)	0.572	0.371	65.78	53.09	25.34	63.47	0.405	0.559	74.49	67.99	40.10	74.43
DLF	0.649	0.018	46.52	28.62	16.85	34.44	0.374	0.575	73.46	65.90	45.20	73.51
MoMKE	0.442	0.620	78.60	64.33	32.87	78.90	0.385	0.659	77.72	66.83	40.97	77.84
TMDC	0.442	0.622	80.17	64.37	32.78	79.94	0.375	0.683	78.49	67.20	40.66	78.55
RECAP	0.544	0.326	70.50	55.54	33.04	68.55	0.460	0.387	68.01	51.94	34.91	67.55
PVLM	0.493	0.491	74.62	59.34	29.10	74.48	0.404	0.596	74.70	64.66	39.44	74.77
VERO(LLaVA-7B)	0.457	0.562	76.81	64.51	34.75	76.89	0.380	0.678	79.29	69.48	42.13	79.20
VERO(LLaVA-13B)	0.451	0.573	77.77	64.73	35.80	77.70	0.377	0.682	78.84	69.05	40.81	78.76
InstructBLIP(xl)	0.523	0.426	80.31	42.45	24.07	80.49	0.428	0.595	77.66	53.29	31.62	77.09
InstructBLIP(xxl)	0.445	0.592	79.65	60.61	34.35	80.16	0.389	0.627	77.85	58.99	37.04	77.99
InstructBLIP(reg xl)	0.418	0.644	79.82	69.02	39.69	79.90	0.457	0.498	70.50	58.76	34.68	70.34
CoRe(xl)	0.375	0.695	82.54	70.46	43.33	82.58	0.314	0.760	83.09	75.20	**48.72**	83.11
CoRe(xxl)	**0.352**	**0.721**	**82.89**	**72.39**	**48.32**	**82.88**	**0.311**	**0.772**	**84.14**	**76.31**	47.78	**84.15**

Table 3: Ablation study on CMU-MOSI and CH-SIMS v2. Acc and F1 are reported in %. For CMU-MOSI, Acc_2 and F1 are reported under the negative/positive setting.

Variant	CMU-MOSI						CH-SIMS v2					
	MAE ↓	Corr ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	F1 ↑	MAE ↓	Corr ↑	Acc_2 ↑	Acc_3 ↑	Acc_5 ↑	F1 ↑
w/o CDS (Random)	0.651	0.849	87.68	53.06	46.09	87.63	0.324	0.742	81.51	74.27	47.33	81.52
w/o C2F (Direct)	0.733	0.806	84.39	48.69	42.39	84.31	0.373	0.661	77.80	69.19	41.80	77.91
w/o Soft Expectation (Hard)	0.749	0.796	82.68	53.03	44.37	82.01	0.406	0.657	80.39	73.68	42.82	80.38
w/o W_1 (Hard CE)	0.647	0.844	87.29	53.21	46.01	87.27	0.332	0.732	81.59	73.89	44.20	81.61
w/o W_1 (MAE)	0.637	0.856	88.54	53.32	45.83	**88.56**	0.338	0.716	80.15	72.65	45.16	80.25
w/o W_1 (Soft CE)	0.632	**0.859**	88.23	54.23	46.73	88.23	0.329	0.741	82.09	74.10	45.03	82.10
w/o Graph (MLP)	0.626	0.854	88.35	54.23	47.03	88.36	0.323	0.743	81.43	73.68	46.91	81.49
w/o RelationGate (Similarity)	0.631	0.852	88.17	53.97	46.91	88.18	0.338	0.722	81.20	73.21	45.74	81.18
w/o Context (Target-only)	0.622	0.853	88.54	54.58	47.70	88.53	0.320	0.754	82.61	75.13	48.36	82.68
CoRe(xl)	**0.619**	0.855	**88.57**	**55.45**	**48.25**	**88.56**	**0.314**	**0.760**	**83.09**	**75.20**	**48.72**	**83.11**

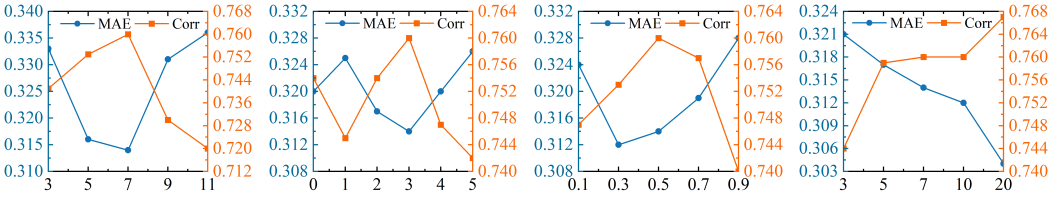
4.2 Ablation Analysis

Table 3 reports the ablation results based on InstructBLIP with Flan-T5-XL. w/o CDS (Random) replaces continuous distribution sampling with random sampling; w/o C2F (Direct) removes coarse-to-fine prediction and performs direct regression; and w/o Soft Expectation (Hard) uses the highest-probability semantic anchor as the prediction. For the training objective, w/o W_1 (MAE), w/o W_1 (Hard CE), and w/o W_1 (Soft CE) replace the original objective with MAE, MAE + Hard CE, and MAE + Soft CE, respectively. For structural modeling, w/o Graph (MLP) replaces the heterogeneous graph with an MLP, w/o RelationGate (Similarity) uses similarity alone for relation modeling, and w/o Context (Target-only) retains only the target sample. Overall, the ablations generally degrade the primary regression performance, with the coarse-to-fine variants showing the largest drops. Both direct regression and hard-anchor prediction substantially worsen MAE and Corr, demonstrating the importance of softly transforming coarse-grained affective semantics into continuous predictions. Random sampling also consistently degrades regression performance, supporting the effectiveness of distribution-preserving sampling under limited supervision. Replacing the original objective with alternative regression or classification losses generally leads to inferior regression performance, supporting the effectiveness of distance-aware supervision. The structural variants also generally degrade performance, validating the heterogeneous graph and relation gate, while contextual information provides additional but relatively modest gains. The full CoRe achieves the best MAE on both datasets and the best Corr on CH-SIMS v2, while remaining highly competitive on the discretized evaluation metrics. Further ablation analyses are provided in Appendices A.3 and A.4.



(a) Label distribution (b) Cumulative distribution (c) Distribution distance

Figure 3: Comparison of label distributions between the full training set and the few-shot subsets obtained by random sampling and continuous distribution sampling on CMU-MOSI.



(a) Anchors (b) Context (c) Loss weight λ (d) Few-shot ratio (%)

Figure 4: Sensitivity analysis of key hyperparameters of CoRe on CH-SIMS v2.

4.3 Effectiveness of Continuous Distribution Sampling

To further evaluate the effectiveness of Continuous Distribution Sampling (CDS), we compare the ground-truth label distribution of the full training set with those of the few-shot subsets obtained by random sampling and CDS on CMU-MOSI from three complementary perspectives. As shown in Figure 3(a), the KDE-based label density of the CDS subset more closely follows that of the full training set, while random sampling exhibits a larger distributional deviation. Figure 3(b) shows a similar trend in the empirical cumulative distribution, where the CDS subset better matches the full training set across the sentiment range. Figure 3(c) further provides a quantitative comparison using Wasserstein-1 distance, where CDS reduces the distance to the full training set. These results show that CDS more effectively preserves the original continuous sentiment distribution under a limited data budget, yielding a more representative few-shot training set. More results are in Appendix A.5.

4.4 Parameter Sensitivity Analysis

We further investigate the sensitivity of CoRe to four key hyperparameters on CH-SIMS v2: the number of anchors, the context size, the loss weight λ, and the few-shot training ratio, as shown in Figure 4. Overall, CoRe remains relatively stable across different settings. A moderate number of anchors yields better performance, with both overly sparse and overly dense anchor configurations causing degradation. The context size shows a similar trend, with moderate contextual information providing the best performance. For λ, intermediate values achieve a better balance between pointwise regression and distance-aware supervision. As the few-shot training ratio increases, MAE consistently decreases, while Corr generally improves and gradually saturates, confirming the benefit of additional supervision. Further sensitivity analysis on CMU-MOSI is provided in Appendix A.6.

5 Conclusion

This paper investigates few-shot MSR and proposes CoRe, a coarse-to-fine learning framework. CoRe constructs a representative few-shot training set that approximates the global sentiment distribution, refines pretrained affective semantics into continuous sentiment predictions, captures cross-modal and contextual dependencies via a heterogeneous multimodal graph, and introduces distance-aware distributional supervision. Comprehensive experiments on four benchmark datasets demonstrate that CoRe consistently outperforms state-of-the-art baselines across multiple evaluation metrics. These results confirm the effectiveness of CoRe and highlight the potential of coarse-to-fine learning for continuous sentiment regression under limited supervision.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: assisting with translation and designing or providing feedback on experiments. We did not use generative AI tools for the following other tasks with required disclosure: helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, implementing methods, supporting qualitative and thematic data analysis, or interpreting results. In addition, the following remaining tasks with required disclosure are not applicable to this work: generating synthetic data sets and cleaning and reformatting datasets. We also used generative AI tools for the following task with recommended disclosure: editing the research paper to improve readability. We have reviewed all AI-assisted work. Specifically, we manually checked the terminology and wording of AI-assisted translations and confirmed their accuracy; any AI feedback on experiments was treated only as reference, with final decisions made by the authors; and AI-assisted editing was used only to improve readability, with all changes reviewed by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work. The anonymized source code is provided as part of the supplementary material for double-blind review. Section 4 describes the experimental setup, including datasets, model configurations, training details, and evaluation metrics. Additional dataset statistics and implementation details, including the prompt and semantic anchor configurations, are provided in Appendices A.1 and A.7. These materials provide the information necessary to reproduce the main results of this work.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Jaime S. Cardoso, Ricardo P. M. Cruz, and Tomé Albuquerque. Unimodal distributions for ordinal regression. *IEEE Trans. Artif. Intell.*, 6(9):2498–2509, 2025.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems, New Orleans, LA, USA*, pp. 49250–49267, 2023.

Yao Du, Qiang Zhai, Weihang Dai, and Xiaomeng Li. Teach CLIP to develop a number sense for ordinal regression. In *Proceedings of the European Conference on Computer Vision, Milan, Italy*, volume 15143, pp. 1–17. Springer, 2024.

Zirun Guo, Tao Jin, Wenlong Xu, Wang Lin, and Yangyang Wu. Bridging the gap for test-time multimodal sentiment analysis. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA*, pp. 16987–16995. AAAI Press, 2025.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Virtual Event / Punta Cana, Dominican Republic*, pp. 9180–9192. Association for Computational Linguistics, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA*, pp. 1122–1131. ACM, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: decoding-enhanced Bert with disentangled attention. In *Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria*. OpenReview.net, 2021.

Chenfeng Huang and George Michailidis. PAC-Bayesian meta-learning for few-shot identification of linear dynamical systems. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Huiting Huang, Tieliang Gong, Kai He, Wen Wen, Weizhan Zhang, and Mengling Feng. Recovering coherent affective patterns: Addressing modality missing in multimodal sentiment analysis. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence, Singapore*, pp. 21957–21965. AAAI Press, 2026.

Tomoharu Iwata, Atsutoshi Kumagai, and Yasutoshi Ida. Meta-learning task-specific regularization weights for few-shot linear regression. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics, Mai Khao, Thailand*, volume 258, pp. 3052–3060. PMLR, 2025.

Ziyu Jia, Tingyu Du, Zhengyu Tian, Hongkai Li, Yong Zhang, and Chenyu Liu. A multimodal Bi-Mamba network with test-time adaptation for emotion recognition based on physiological signals. In *Advances in Neural Information Processing Systems, San Diego, CA, USA*, 2025.

Ran Jing, Geng Tu, Yice Zhang, and Ruifeng Xu. Causal-ERC: A multimodal framework with causal prompting for emotion recognition in conversations with large language models. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence, Singapore*, pp. 31383–31391. AAAI Press, 2026.

Byeongjin Jung, Chanyeong Park, and Sejoon Lim. Distance-aware soft prompt guidance for multimodal valence-arousal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5294–5301, June 2026.

Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. OrdinalCLIP: Learning rank prompts for language-guided ordinal regression. In *Advances in Neural Information Processing Systems, New Orleans, LA, USA*, pp. 35313–35325, 2022.

Zhaoying Li, Zhaohu Xing, Hongying Liu, Lei Zhu, and Liang Wan. Anchored supervised contrastive learning for long-tailed medical image regression. In *Pattern Recognition and Computer Vision, Urumqi, China*, volume 15045 of *Lecture Notes in Computer Science*, pp. 3–18. Springer, 2024.

Dongjun Liu, Weichen Dai, Hangjie Yi, Honggang Liu, Jianting Cao, Qibin Zhao, Fabio Babiloni, and Wanzeng Kong. Brain-machine enhanced intelligence for semi-supervised facial emotion recognition. *IEEE Trans. Affect. Comput.*, 17(1):119–131, 2026.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the International Conference on Multimodal Interaction, Bengaluru, India*, pp. 247–258. ACM, 2022.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia*, pp. 2247–2256. Association for Computational Linguistics, 2018.

Miaosen Luo, Yuncheng Jiang, and Sijie Mai. Towards explainable fusion and balanced learning in multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland*, pp. 1997–2006. ACM, 2025.

Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.*, 2(2): 92–105, 2011.

Guangkun Nie, Gongzheng Tang, and Shenda Hong. Dist loss: Enhancing regression in few-shot region through distribution distance constraint. In *Proceedings of the Thirteenth International Conference on Learning Representations, Singapore*. OpenReview.net, 2025.

Georgios Paltoglou and Mike Thelwall. Seeing stars of valence and arousal in blog posts. *IEEE Trans. Affect. Comput.*, 4(1):116–123, 2013.

Xihang Qiu, Jiarong Cheng, Yuhao Fang, Wanpeng Zhang, Yao Lu, Ye Zhang, and Chun Li. Federated dialogue-semantic diffusion for emotion recognition under incomplete modalities. In *Advances in Neural Information Processing Systems, San Diego, CA, USA*, pp. 86146–86175, 2025.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web*, pp. 593–607, Cham, 2018. Springer International Publishing.

Marcin Sendera, Jacek Tabor, Aleksandra Nowak, Andrzej Bedychaj, Massimiliano Patacchiola, Tomasz Trzcinski, Przemyslaw Spurek, and Maciej Zieba. Non-Gaussian Gaussian processes for few-shot regression. In *Advances in Neural Information Processing Systems, virtual*, pp. 10285–10298, 2021.

Peng Shi, Guoyan Huang, Hongdou He, Guyu Zhao, Xiaobing Hao, and Yifang Huang. Few-shot regression with differentiable reference model. *Inf. Sci.*, 658:120010, 2024.

Kai Sun, Hao Wu, Bin Shi, Samuel Mensah, Peng Liu, and Bo Dong. VERO: verification and zero-shot feedback acquisition for few-shot multimodal aspect-level sentiment classification. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA*, pp. 25210–25218. AAAI Press, 2025.

Mani Kumar Tellamekala, Shahin Amiriparian, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):805–822, 2024.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*, pp. 6558–6569. Association for Computational Linguistics, 2019a.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *Proceedings of the International Conference on Learning Representations*, 2019b.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, Lihuo He, and Xuemei Luo. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognit.*, 136: 109259, 2023a.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Trans. Multim.*, 25:4909–4921, 2023b.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA*, pp. 21180–21188. AAAI Press, 2025.

Rui Wang, Peipei Li, Huaibo Huang, Chunshui Cao, Ran He, and Zhaofeng He. Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. In *Advances in Neural Information Processing Systems, New Orleans, LA, USA*, pp. 76908–76922, 2023c.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language*, pp. 11–19, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia*, pp. 438–446. ACM, 2024.

Xingle Xu, YongKang Liu, Dexian Cai, Shi Feng, Xiaocui Yang, Daling Wang, and Yifei Zhang. MoLAN: A unified modality-aware noise dynamic editing framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics*, pp. 24480–24495, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Xiaocui Yang, Shi Feng, Daling Wang, Qi Sun, Wenfang Wu, Yifei Zhang, Pengfei Hong, and Soujanya Poria. Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. In *Findings of the Association for Computational Linguistics, Toronto, Canada*, pp. 11575–11589. Association for Computational Linguistics, 2023a.

Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Soujanya Poria. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. In *Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada*, pp. 6045–6053. ACM, 2023b.

Yang Yang, Xunde Dong, and Yupeng Qiang. MSE-Adapter: A lightweight plugin endowing LLMs with the capability to perform multimodal sentiment analysis and emotion recognition. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, pp. 25642–25650, Philadelphia, PA, USA, February 2025. AAAI Press.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727. Association for Computational Linguistics, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual Event*, pp. 10790–10797. AAAI Press, 2021.

Yang Yu and Dong Zhang. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In *IEEE International Conference on Multimedia and Expo, Taipei, Taiwan*, pp. 1–6. IEEE, 2022.

Yang Yu, Dong Zhang, and Shoushan Li. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal*, pp. 189–198. ACM, 2022.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.*, 31(6):82–88, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark*, pp. 1103–1114. Association for Computational Linguistics, 2017.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA*, pp. 5634–5641. AAAI Press, 2018a.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia*, pp. 2236–2246. Association for Computational Linguistics, 2018b.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 756–767, Singapore, December 2023a. Association for Computational Linguistics.

Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep regression with ordinal entropy. In *Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda*. OpenReview.net, 2023b.

Chen Zhao and Feng Chen. Unfairness discovery and prevention for few-shot regression. In *2020 IEEE International Conference on Knowledge Graph, Online*, pp. 137–144. IEEE, 2020.

Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021.

Li Zheng, Yanyi Luo, Hao Fei, Yuzhe Ding, Yujie Huang, Fei Li, Chong Teng, and Donghong Ji. Dynamic emotion and personality profiling for multimodal deception detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, San Diego, California, United States*, pp. 3930–3940. Association for Computational Linguistics, 2026.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. GLoMo: Global-local modal fusion for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia*, pp. 1800–1809. ACM, 2024.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. TMDC: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence, Singapore*, pp. 2281–2289. AAAI Press, 2026a.

Yan Zhuang, Yanru Zhang, Jiawen Deng, and Fuji Ren. Intra-sample and intra-modal enhancement for multimodal sentiment analysis with missing modalities. *IEEE Trans. Multim.*, 28:1847–1859, 2026b.

Table 4: Statistics of the four benchmark datasets and their corresponding few-shot training subsets. V and S denote the numbers of videos and segments, respectively, while "Var." indicates that the number of retained segments varies across random seeds.

Dataset	Train		Validation		Test		Few-shot Training (7%)	
	V	S	V	S	V	S	V	S
CMU-MOSI	52	1,284	10	229	31	686	4	Var.
CMU-MOSEI	2,216	16,326	296	1,871	670	4,658	156	Var.
CH-SIMS	60	1,368	59	456	59	457	–	96
CH-SIMS v2	134	2,722	120	647	145	1,034	–	191

Table 5: Standard deviations on CMU-MOSI and CMU-MOSEI. Acc and F1 are reported in %.

Method	CMU-MOSI						CMU-MOSEI					
	MAE	Corr	Acc_2	Acc_5	Acc_7	F1	MAE	Corr	Acc_2	Acc_5	Acc_7	F1
TFN	0.064	0.024	2.13/2.36	2.31	1.60	2.29/2.52	0.009	0.025	1.76/1.21	0.56	0.51	1.40/1.25
TETFN	0.021	0.079	4.60/6.42	0.00	0.00	9.52/10.82	0.006	0.006	0.95/0.27	0.75	0.72	0.78/0.31
Self-MM	0.047	0.019	1.58/1.77	1.37	1.29	1.60/1.78	0.002	0.005	1.13/0.33	0.16	0.13	0.93/0.28
MulT	0.092	0.070	8.16/9.52	6.24	4.94	13.08/14.44	0.017	0.041	1.22/1.74	0.95	0.75	1.27/2.00
MMIM	0.062	0.022	1.69/1.85	3.48	2.74	1.75/1.90	0.005	0.004	1.73/0.48	1.05	0.84	1.37/0.52
MISA	0.053	0.028	1.72/1.98	1.96	1.33	1.81/2.05	0.006	0.007	0.63/0.37	0.66	0.53	0.50/0.33
MFN	0.066	0.036	3.12/3.42	2.96	2.27	3.38/3.68	0.008	0.011	1.61/1.46	0.80	0.57	1.45/1.40
MFM	0.043	0.016	1.12/1.39	1.85	1.19	1.26/1.52	0.021	0.070	4.16/3.27	0.93	1.02	3.80/3.15
LMF	0.111	0.259	9.23/9.35	6.44	5.21	9.17/9.26	0.006	0.016	1.91/1.08	0.24	0.25	1.60/1.11
LF-DNN	0.012	0.005	1.65/1.72	1.26	1.29	1.70/1.78	0.007	0.009	2.10/0.97	0.65	0.59	1.69/0.75
Graph-MFN	0.047	0.015	1.31/1.52	2.28	1.31	1.47/1.67	0.008	0.011	0.96/0.66	0.46	0.39	0.82/0.65
EF-LSTM	0.099	0.058	10.59/12.11	5.69	5.01	17.35/18.93	0.012	0.030	1.22/1.76	0.55	0.50	1.38/2.39
CENET	0.053	0.047	2.00/2.10	2.07	1.77	2.02/2.11	0.014	0.015	0.72/0.67	0.94	0.79	0.64/0.62
ALMT	0.023	0.019	1.10/1.23	1.13	0.48	1.10/1.23	0.005	0.004	0.85/0.33	0.56	0.51	0.73/0.34
MSE-Adapter	0.037	0.017	1.05/1.44	1.94	1.40	1.04/1.53	0.020	0.033	0.47/6.37	0.78	0.84	0.72/7.05
KAN-MCP	0.099	0.104	5.21/5.73	3.44	3.29	5.45/5.95	0.015	0.113	0.66/2.49	0.00	0.00	4.49/6.89
GLoMo	0.028	0.052	3.02/3.17	3.33	3.33	3.10/3.25	0.006	0.006	1.02/0.53	0.60	0.55	0.80/0.48
IIE	0.049	0.034	2.03/2.14	2.21	1.69	2.07/2.18	0.003	0.003	1.11/0.61	0.31	0.34	0.97/0.57
DLF	0.038	0.058	5.01/7.43	0.00	0.00	6.84/9.47	0.017	0.025	1.42/1.35	1.04	0.88	1.31/1.31
MoMKE	0.029	0.013	1.59/1.89	1.09	0.40	1.70/1.98	0.009	0.010	1.40/0.20	0.73	0.80	1.16/0.17
TMDC	0.011	0.008	0.77/0.93	2.00	1.58	0.75/0.89	0.036	0.006	0.61/0.46	1.22	1.92	0.54/0.46
RECAP	0.051	0.050	2.02/2.17	0.95	1.18	2.25/2.68	0.033	0.157	1.24/4.20	0.38	0.25	5.45/10.34
PVLM	0.027	0.008	1.06/1.15	2.16	1.90	1.07/1.15	0.004	0.006	0.91/0.58	0.41	0.39	0.83/0.54
VERO(7B)	0.026	0.010	0.71/0.67	2.05	1.66	0.73/0.65	0.019	0.019	1.23/1.00	1.90	1.63	1.11/0.94
VERO(13B)	0.014	0.007	0.55/0.40	1.25	1.17	0.59/0.42	0.014	0.011	2.24/0.64	1.38	1.24	1.93/0.57
InstructBLIP(reg xl)	0.011	0.008	0.66/0.69	0.85	0.65	0.66/0.68	0.005	0.005	1.36/0.71	0.97	0.65	1.23/0.67
CoRe(xl)	0.008	0.004	0.49/0.39	1.53	1.33	0.51/0.39	0.004	0.003	2.26/0.82	0.43	0.48	2.00/0.73
CoRe(xxl)	0.002	0.001	0.40/0.29	0.45	0.40	0.42/0.30	0.002	0.003	0.43/0.10	0.50	0.33	0.36/0.09

A Appendix

A.1 Dataset Statistics

Table 4 summarizes the standard training, validation, and test splits of the four benchmark datasets, together with the few-shot training subsets constructed under a 7% sampling ratio. For CMU-MOSI and CMU-MOSEI, the few-shot subsets are constructed by sampling complete videos. Since different random seeds may select different videos, the number of retained segments varies across runs. For CH-SIMS and CH-SIMS v2, individual segments are sampled directly, and the subset size is determined by the specified sampling ratio.

A.2 Stability Analysis

Tables 5 and 6 report the standard deviations over five independent runs, corresponding to the mean results in Tables 1 and 2. Overall, CoRe exhibits stable performance across different random seeds, with small variations in the primary regression metrics across all four datasets. The discretized metrics also remain generally stable across runs. Compared with several baselines that exhibit substantially larger cross-run variations under the few-shot setting, CoRe maintains more consistent performance across datasets and model scales. These results demonstrate the robustness of CoRe to different random seeds.

Table 6: Standard deviations on CH-SIMS and CH-SIMS v2. Acc and F1 are reported in %.

Method	CH-SIMS						CH-SIMS v2					
	MAE	Corr	Acc_2	Acc_3	Acc_5	F1	MAE	Corr	Acc_2	Acc_3	Acc_5	F1
TFN	0.014	0.040	1.86	2.92	2.36	1.36	0.010	0.031	1.29	1.80	1.84	1.28
TETFN	0.011	0.050	18.19	13.10	1.51	17.13	0.005	0.117	5.46	11.83	6.75	7.74
Self-MM	0.035	0.059	17.39	11.71	2.65	19.57	0.010	0.020	1.30	1.38	0.78	1.29
MulT	0.070	0.127	12.25	8.06	3.30	14.20	0.010	0.019	1.07	0.85	0.46	1.02
MMIM	0.026	0.059	1.67	2.22	2.88	2.15	0.006	0.014	0.60	0.81	0.85	0.58
MISA	0.017	0.040	2.82	2.45	2.38	2.76	0.049	0.163	6.17	11.17	7.10	6.40
MFN	0.036	0.129	3.93	5.43	3.98	5.25	0.029	0.071	2.27	3.37	2.32	2.54
MFM	0.005	0.078	0.00	0.10	0.10	0.00	0.005	0.024	6.19	15.78	9.48	6.86
LMF	0.032	0.164	5.95	10.03	2.91	6.13	0.011	0.033	1.24	1.18	1.52	1.08
LF-DNN	0.019	0.047	2.62	2.04	3.48	2.38	0.026	0.047	2.08	2.76	2.30	2.04
Graph-MFN	0.024	0.068	3.24	2.26	2.39	3.20	0.043	0.201	5.89	6.10	3.84	11.45
EF-LSTM	0.002	0.044	0.00	19.19	3.00	0.00	0.028	0.061	1.93	1.91	0.85	1.95
CENET	0.007	0.059	2.55	2.73	1.15	1.32	0.010	0.006	0.48	0.35	0.53	0.46
ALMT	0.012	0.027	2.71	2.73	1.60	1.99	0.020	0.034	2.48	2.28	2.71	2.46
MSE-Adapter	0.098	0.075	12.58	13.52	5.75	12.90	0.024	0.036	2.03	3.53	3.39	2.11
DLF	0.067	0.071	17.44	13.85	2.27	18.92	0.008	0.019	1.01	0.89	1.23	1.04
MoMKE	0.026	0.041	3.26	1.33	1.70	2.88	0.009	0.015	0.78	1.34	1.62	0.74
TMDC	0.033	0.069	2.46	3.16	3.46	2.77	0.017	0.021	0.96	0.92	1.74	1.00
RECAP	0.017	0.057	0.77	0.47	0.41	0.78	0.020	0.018	1.70	3.93	2.16	1.64
PVLM	0.020	0.056	3.10	2.40	1.81	3.00	0.003	0.006	0.78	0.68	0.65	0.75
VERO(7B)	0.032	0.066	2.35	2.96	4.00	2.23	0.012	0.017	0.99	1.48	0.73	1.02
VERO(13B)	0.025	0.052	2.72	3.52	4.27	2.71	0.010	0.019	1.56	1.74	0.96	1.51
InstructBLIP(reg xl)	0.007	0.008	1.33	1.46	3.09	1.26	0.017	0.025	1.59	1.52	1.58	1.92
CoRe(xl)	0.014	0.021	1.33	2.12	3.38	1.16	0.009	0.009	0.82	0.85	2.30	0.80
CoRe(xxl)	0.021	0.027	1.53	1.46	4.33	1.25	0.007	0.011	1.09	1.47	2.27	0.79

Table 7: Ablation study of CDS on CMU-MOSI and CH-SIMS v2. Acc and F1 are reported in %. For CMU-MOSI, Acc_2 and F1 are reported under the negative/positive setting.

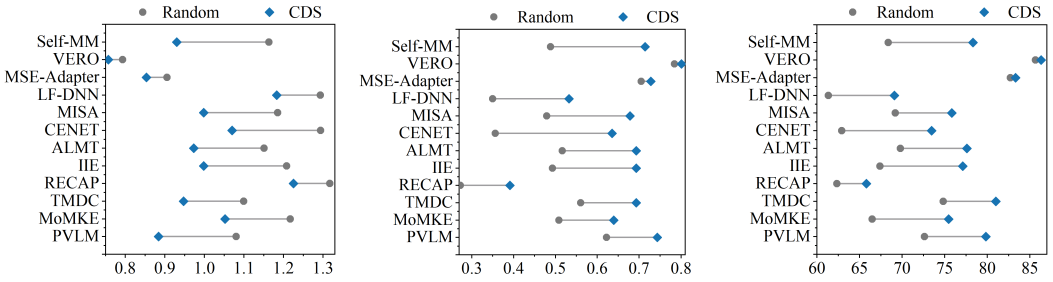
Variant	CMU-MOSI						CH-SIMS v2					
	MAE ↓	Corr ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	F1 ↑	MAE ↓	Corr ↑	Acc_2 ↑	Acc_3 ↑	Acc_5 ↑	F1 ↑
w/ Uniform	0.673	0.849	87.18	51.81	44.90	87.17	0.342	0.718	80.08	72.31	44.26	80.07
w/o KDE	0.624	**0.862**	87.70	54.91	47.52	87.73	0.334	0.729	80.95	73.54	45.76	80.99
w/ L1	0.643	0.850	87.65	55.35	47.86	87.61	–	–	–	–	–	–
w/ JS	0.659	0.840	87.77	54.52	47.26	87.77	–	–	–	–	–	–
CoRe(xl)	**0.619**	0.855	**88.57**	**55.45**	**48.25**	**88.56**	**0.314**	**0.760**	**83.09**	**75.20**	**48.72**	**83.11**

A.3 Ablation Analysis of Continuous Distribution Sampling

Table 7 further examines the main design choices of CDS. w/ Uniform allocates samples equally across sentiment intervals; w/o KDE removes kernel density estimation and uses empirical interval frequencies. For video-level sampling, w/ L1 and w/ JS replace Wasserstein-1 distance with L1 and Jensen–Shannon divergence, respectively. Overall, the original CDS achieves the strongest performance across most metrics on both datasets. Uniform sampling causes the largest degradation, indicating that equal coverage of the sentiment space cannot adequately approximate the underlying sentiment distribution. Removing KDE also generally degrades performance, despite a slightly higher Corr on CMU-MOSI, supporting the benefit of continuous density estimation. On CMU-MOSI, replacing Wasserstein-1 with L1 or JS divergence yields inferior overall performance, demonstrating the advantage of accounting for the distances between sentiment intervals during distribution matching.

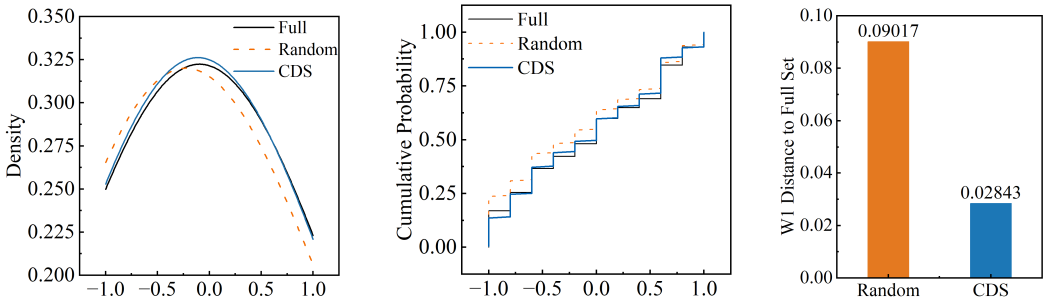
A.4 Effect of Continuous Distribution Sampling Across Baselines

To further examine whether the benefit of CDS extends across different models, we compare random sampling and CDS on 12 multimodal baselines under the same 7% few-shot sampling ratio, while keeping all other settings unchanged for each baseline. As shown in Figure 5, CDS consistently improves performance across diverse baselines, yielding lower MAE and higher Corr and Acc_2. The improvements are observed across models with substantially different architectures and performance



(a) MAE ↓ (b) Corr ↑ (c) Acc_2 (%) ↑

Figure 5: Performance comparison of 12 multimodal baselines using random sampling and continuous distribution sampling under the few-shot setting on CMU-MOSI.



(a) Label distribution (b) Cumulative distribution (c) Distribution distance

Figure 6: Comparison of the full training set with the few-shot subsets obtained by random sampling and continuous distribution sampling on CH-SIMS v2.

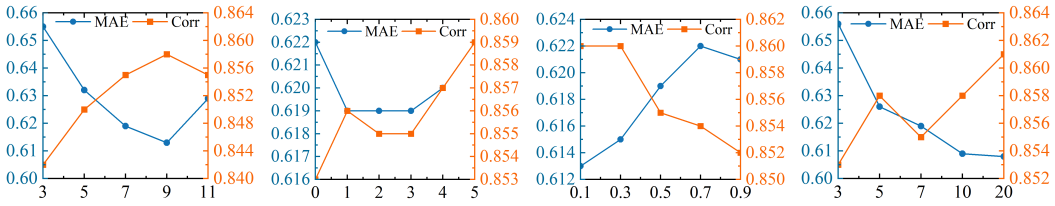
levels, indicating that the effectiveness of CDS is not specific to CoRe or a particular baseline. These results demonstrate that constructing a more representative few-shot training subset can generally benefit multimodal sentiment regression under limited supervision.

A.5 Distribution Preservation on CH-SIMS v2

We further examine the distribution-preserving ability of CDS on CH-SIMS v2. Figure 6(a) compares the label distributions estimated by KDE, where the CDS subset more closely follows the full training set than the randomly sampled subset. A similar pattern is observed in the empirical cumulative distributions in Figure 6(b), with CDS showing closer agreement with the full training set across the sentiment range. Figure 6(c) provides a quantitative comparison using Wasserstein-1 distance, which decreases from 0.09017 for random sampling to 0.02843 for CDS. These results further demonstrate that CDS effectively preserves the original continuous sentiment distribution under a limited data budget.

A.6 Additional Sensitivity Analysis on CMU-MOSI

To further evaluate the robustness of CoRe across datasets, we conduct additional sensitivity analysis on CMU-MOSI, as shown in Figure 7. Overall, CoRe remains relatively stable across different hyperparameter settings. Performance varies only mildly for moderate anchor numbers, with only minor differences between seven and nine anchors, while too few or too many anchors lead to degradation. Varying the context size also causes only minor fluctuations, indicating limited sensitivity to contextual settings. For the loss weight λ, relatively small values generally yield better performance, with $\lambda = 0.1$ achieving the lowest MAE, while larger values lead to degradation. As the few-shot training ratio increases, MAE generally decreases, and Corr improves overall, further demonstrating the benefit of additional supervision.



(a) Anchors (b) Context (c) Loss weight λ (d) Few-shot ratio (%)

Figure 7: Sensitivity analysis of key hyperparameters of CoRe on CMU-MOSI.

Table 8: Semantic anchor configurations for different datasets.

Dataset	Anchor values	Semantic descriptions
CMU-MOSI / CMU-MOSEI	$-3, -2, -1, 0, 1, 2, 3$	strongly negative, negative, weakly negative, neutral, weakly positive, positive, strongly positive
CH-SIMS / CH-SIMS v2	$-1, -\frac{2}{3}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{2}{3}, 1$	strongly negative, moderately negative, slightly negative, neutral, slightly positive, moderately positive, strongly positive

A.7 Prompt Configuration

To ensure reproducibility, we summarize the semantic anchor and prompt configurations used in our experiments. The anchor values and their corresponding semantic descriptions for each dataset are listed in Table 8. All prompt templates follow the same structure, consisting of a multimodal sentiment instruction, the input utterance, the dataset-specific semantic anchor descriptions, and a label-only response constraint. The common prompt form is:

```
Given the multimodal content, assign a sentiment label from
{anchors}.
Return label only without any other text.
Sentence: {text}
Label:
```

Here, {text} denotes the input utterance, while {anchors} denotes the dataset-specific semantic descriptions listed in Table 8.

A.8 Efficiency Analysis

To evaluate the computational overhead of CoRe, we compare total and trainable parameters, inference latency and memory, and training latency and memory under the same hardware environment and batch size. As shown in Table 9, CoRe-XL and CoRe-XXL require only 4.60M and 5.65M trainable parameters and achieve inference latencies of 169.28 and 181.08 ms/sample, respectively, both lower than VERO-7B, VERO-13B, and MSE-Adapter-6B. Their peak inference memory is 8.17 and 23.85 GiB, respectively. Frozen-XL and Frozen-XXL provide efficiency references using the same pretrained backbones with scalar regression heads. Compared with these references, CoRe introduces additional computation for heterogeneous graph modeling, structural modulation, and coarse-to-fine prediction, resulting in increased training cost. Nevertheless, the inference memory overhead remains small, increasing from 7.91 to 8.17 GiB for XL and from 23.36 to 23.85 GiB for XXL. Moreover, CoRe maintains competitive training and inference efficiency compared with the other baselines. Together with the performance improvements in Tables 1 and 2, these results demonstrate that CoRe improves few-shot regression performance with limited parameter and memory overhead while maintaining competitive computational efficiency.

A.9 Comparison with Fully Supervised Methods

To further evaluate the effectiveness of CoRe under limited supervision, we compare CoRe trained with only 7% of the training data against representative fully supervised methods trained on the complete training set. As shown in Tables 10 and 11, CoRe remains highly competitive despite using substantially fewer labeled samples. On CMU-MOSI, CoRe achieves the best MAE and Corr

Table 9: Computational efficiency comparison across different methods.

Method	Total Params	Trainable Params	Inference Latency (ms/sample)	Inference Memory (GiB)	Training Latency (ms/step)	Training Memory (GiB)
Frozen-XL	4.02B	1.05M	112.57	7.91	96.70	7.93
Frozen-XXL	12.31B	2.10M	106.97	23.36	108.85	23.38
VERO-7B	7.07B	4.20M	216.00	13.43	404.12	16.74
VERO-13B	13.36B	6.56M	370.46	25.39	769.16	30.54
MSE-Adapter-6B	6.25B	2.61M	320.67	12.18	158.41	23.47
CoRe-XL	4.03B	4.60M	169.28	8.17	321.00	8.96
CoRe-XXL	12.32B	5.65M	181.08	23.85	335.27	25.40

Table 10: Comparison with fully supervised methods on CMU-MOSI and CMU-MOSEI. Baselines are trained on the full training set (100%), while CoRe uses only 7% of the training data.

Method	Ratio	CMU-MOSI						CMU-MOSEI					
		MAE ↓	Corr ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	F1 ↑	MAE ↓	Corr ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	F1 ↑
MSE-Adapter	100%	0.801	0.753	82.36/81.55	48.83	41.40	82.38/81.31	0.515	0.783	86.91/-	–	54.56	86.77/-
TMDC	100%	0.734	0.815	86.89/88.26	44.61	39.36	86.83/88.15	0.578	0.799	84.11/87.23	52.72	49.45	84.11/87.20
DLF	100%	0.731	0.781	-/85.06	52.33	47.08	-/85.04	0.536	0.764	-/85.42	55.70	53.90	-/85.27
KAN-MCP	100%	0.615	0.857	-/89.40	–	48.30	-/89.40	0.522	0.788	-/87.70	–	53.90	-/87.60
RECAP	100%	0.723	0.791	83.82/85.52	52.19	46.79	83.90/85.59	0.551	0.764	83.30/85.50	54.56	52.39	83.12/85.63
ALMT	100%	0.736	0.788	82.36/84.30	50.73	45.92	82.29/84.29	0.547	0.763	79.91/84.73	53.32	51.77	80.58/84.77
CoRe(xxl)	7%	0.558	0.876	87.06/89.33	62.24	52.77	87.01/89.33	0.537	0.780	84.00/87.50	53.99	52.66	84.40/87.49

among the compared methods and performs competitively on the discretized metrics. On CMU-MOSEI, CH-SIMS, and CH-SIMS v2, CoRe also matches or surpasses fully supervised baselines on several regression and classification metrics. These results demonstrate that CoRe can achieve strong continuous sentiment prediction performance with substantially reduced supervision.

A.10 Implementation Details of Continuous Distribution Sampling

We follow the native data organization of each benchmark. CMU-MOSI and CMU-MOSEI adopt video-level sampling to preserve complete video structure, whereas CH-SIMS and CH-SIMS v2 adopt segment-level sampling according to their segment-based organization.

Segment-level sampling. When individual segments are used as sampling units, given a total sampling budget K and the target distribution $\boldsymbol{p} = [p_1, \ldots, p_B]$, the desired number of samples assigned to interval $\mathcal{I}_b$ is Kp_b. Since Kp_b is generally non-integer, we first take its integer part and allocate the remaining budget in descending order of the fractional residuals, ensuring

$$\sum_{b=1}^{B} k_b = K \tag{18}$$

where k_b denotes the final number of segments selected from $\mathcal{I}_b$. We then randomly select k_b segments from each interval and combine them to form the few-shot training set.

Video-level sampling. When complete videos are used as sampling units, each video is represented by the distribution of its segment-level zero-shot predictions over the same B intervals:

$$\boldsymbol{h}_m = [h_{m,1}, \ldots, h_{m,B}] \tag{19}$$

where $h_{m,b}$ denotes the number of segments in the m-th video whose zero-shot predictions fall into interval $\mathcal{I}_b$. For a candidate video subset $\mathcal{S}$, the aggregated histogram is normalized as:

$$q_{\mathcal{S},b} = \frac{\sum_{m\in\mathcal{S}} h_{m,b}}{\sum_{r=1}^{B}\sum_{m\in\mathcal{S}} h_{m,r}}, \qquad \boldsymbol{q}_{\mathcal{S}} = [q_{\mathcal{S},1}, \ldots, q_{\mathcal{S},B}] \tag{20}$$

Given a video budget K_v, subset selection is formulated as:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}\subseteq\mathcal{V},\,|\mathcal{S}|=K_v} W_1\left(\boldsymbol{q}_{\mathcal{S}}, \boldsymbol{p}\right) \tag{21}$$

where $\mathcal{V}$ is the set of candidate videos. Since exhaustive search over all feasible subsets is computationally prohibitive, we employ incremental greedy selection with multiple initializations, followed by one-video swap refinement. The subset with the smallest Wasserstein-1 distance is retained.

Table 11: Comparison with fully supervised methods on CH-SIMS and CH-SIMS v2. Baselines are trained on the full training set (100%), while CoRe uses only 7% of the training data.

Method	Ratio	CH-SIMS						CH-SIMS v2					
		MAE ↓	Corr ↑	Acc_2 ↑	Acc_3 ↑	Acc_5 ↑	F1 ↑	MAE ↓	Corr ↑	Acc_2 ↑	Acc_3 ↑	Acc_5 ↑	F1 ↑
MSE-Adapter	100%	0.606	–	69.37	15.10	15.10	56.82	0.296	0.720	83.77	–	–	83.76
TMDC	100%	0.354	0.725	83.15	70.02	47.70	83.50	0.315	0.746	80.75	72.34	50.39	80.82
DLF	100%	0.440	0.575	78.99	64.55	39.17	79.19	–	–	–	–	–	–
KAN-MCP	100%	–	–	–	–	–	–	0.281	0.742	81.60	75.00	57.30	81.70
RECAP	100%	0.401	0.618	79.21	67.61	45.73	79.04	0.333	–	78.05	70.79	51.26	78.07
ALMT	100%	0.414	0.594	80.31	66.30	42.01	80.31	0.326	0.655	76.60	68.47	50.48	76.67
CoRe(xxl)	7%	0.352	0.721	82.89	72.39	48.32	82.88	0.311	0.772	84.14	76.31	47.78	84.15

Discrete Wasserstein-1 distance. For two probability distributions P and Q defined on a one-dimensional space, the Wasserstein-1 distance can be expressed as:

$$W_1(P,Q) = \int_{-\infty}^{+\infty} |F_P(y) - F_Q(y)|\, dy \tag{22}$$

where $F_P(y)$ and $F_Q(y)$ denote the corresponding cumulative distribution functions.

After partitioning the continuous sentiment space into B intervals, we define $F_{\boldsymbol{p}}(b) = \sum_{r=1}^{b} p_r$ and $F_{\boldsymbol{q}}(b) = \sum_{r=1}^{b} q_r$. The discrete approximation is then

$$W_1(\boldsymbol{p},\boldsymbol{q}) = \sum_{b=1}^{B-1} |F_{\boldsymbol{p}}(b) - F_{\boldsymbol{q}}(b)|\, \Delta_b \tag{23}$$

where Δ_b denotes the distance between the centers of adjacent intervals. This formulation accounts for both probability-mass mismatch and the distance over which the mass is displaced.

A.11 Implementation Details of Heterogeneous Multimodal Graph Modeling

Context construction and relation weighting. For each target sample i, contextual candidates are restricted to samples from the same video to maintain semantic and temporal relevance. Depending on the organization of the benchmark, the candidate pool is constructed either from a local temporal window around the target sample or from the remaining observable samples within the same video. During training, validation, and testing, candidates are restricted to the corresponding data split to prevent information leakage.

Candidates are ranked according to the multimodal similarity s_{ij} defined in the main text, and the top-k candidates are retained as the contextual neighborhood $\mathcal{N}_i$. If fewer than k candidates are available, all available candidates are retained.

For relation weighting, the modality features of each sample are concatenated and projected into a shared relation space:

$$\bar{\boldsymbol{z}}_i = \boldsymbol{W}_p[\boldsymbol{z}_i^t; \boldsymbol{z}_i^a; \boldsymbol{z}_i^v] + \boldsymbol{b}_p \tag{24}$$

The projected target and contextual representations are used by the relation MLP f_{rel} defined in the main text to obtain g_{ij}. The final contextual edge weight is $\omega_{ij} = s_{ij} g_{ij}$, combining the similarity prior with the learned relation gate.

Heterogeneous graph construction. Each target or contextual sample is represented by three modality-specific nodes corresponding to text, audio, and vision. The modality features are independently projected into the graph hidden space:

$$\boldsymbol{h}_i^{m,(0)} = \boldsymbol{W}_m \boldsymbol{z}_i^m + \boldsymbol{b}_m, \qquad m \in \{t,a,v\} \tag{25}$$

The graph contains two relation types. Cross-modal relations connect the three modality nodes within each sample bidirectionally, forming text–audio, text–visual, and audio–visual connections. Contextual relations connect each modality node of the target sample bidirectionally to the three modality nodes of every selected contextual sample. All contextual edges associated with the same

target–context pair (i, j) share the weight ω_{ij}. No contextual edges are introduced between different contextual samples.

For each relation r, its adjacency matrix $\boldsymbol{A}_r$ is symmetrically normalized as:

$$\widetilde{\boldsymbol{A}}_r = \boldsymbol{D}_r^{-1/2} \boldsymbol{A}_r \boldsymbol{D}_r^{-1/2} \tag{26}$$

where $\boldsymbol{D}_r$ is the corresponding degree matrix. The normalized relation-specific adjacency matrices are then used in the graph propagation defined in the main text.

Target-specific graph readout. After L graph layers, only the three modality nodes corresponding to the target sample are used for graph readout. Their representations are concatenated as:

$$\boldsymbol{h}_i^{\mathrm{tar}} = [\boldsymbol{h}_{i,t}^{(L)}; \boldsymbol{h}_{i,a}^{(L)}; \boldsymbol{h}_{i,v}^{(L)}] \tag{27}$$

The target-specific structural representation is obtained through normalization and linear projection:

$$\boldsymbol{g}_i = \mathrm{LN}_{\mathrm{out}} \left(\boldsymbol{W}_g \, \mathrm{LN}_{\mathrm{read}} \left(\boldsymbol{h}_i^{\mathrm{tar}} \right) + \boldsymbol{b}_g \right) \tag{28}$$

This readout retains the contextual and cross-modal information propagated to the target nodes while producing a compact structural representation for subsequent coarse-to-fine sentiment prediction.

A.12 Theoretical Analysis of Soft Expectation

Continuous support and range preservation. Assume the semantic anchors are ordered as $a_1 < a_2 < \cdots < a_C$. Although the prediction is constructed from a finite set of anchors, the soft expectation is not restricted to discrete anchor values. For any $y \in [a_1, a_C]$, there exists an adjacent pair a_r and a_{r+1} such that $a_r \le y \le a_{r+1}$. Consider the probability assignment:

$$\pi_r = \frac{a_{r+1} - y}{a_{r+1} - a_r}, \qquad \pi_{r+1} = \frac{y - a_r}{a_{r+1} - a_r}, \qquad \pi_c = 0, \quad \forall c \notin \{r, r+1\} \tag{29}$$

Since $\pi_r, \pi_{r+1} \ge 0$ and $\pi_r + \pi_{r+1} = 1$, it follows that

$$\pi_r a_r + \pi_{r+1} a_{r+1} = y \tag{30}$$

Therefore,

$$\left\{ \mathbf{a}^\top \boldsymbol{\pi} \mid \boldsymbol{\pi} \in \Delta^{C-1} \right\} = \mathrm{conv}(\mathcal{A}) = [a_1, a_C] \tag{31}$$

where $\mathbf{a} = [a_1, \ldots, a_C]^\top$ and Δ^{C-1} denotes the probability simplex. Thus, soft combinations of a finite set of semantic anchors provide support over the entire continuous sentiment interval rather than imposing hard quantization. Since the prediction is a convex combination of the anchors:

$$a_1 \le \hat{y}_i \le a_C, \tag{32}$$

so the output remains within the valid sentiment range without additional clipping.

Order preservation and differentiability. Consider transferring probability mass δ from anchor a_r to anchor a_s, where $a_s > a_r$ and $0 \le \delta \le \pi_{i,r}$, while keeping all other probabilities unchanged. The resulting change in the prediction is:

$$\hat{y}_i' - \hat{y}_i = \delta(a_s - a_r) \tag{33}$$

Thus, shifting probability mass toward a higher-valued anchor increases the prediction, with the magnitude of the change proportional to the distance between the two anchors. The soft-expectation mapping therefore preserves the numerical ordering of the anchors.

Combining Eqs. (12) and (13), the derivative of the prediction with respect to the NLL of the c-th semantic description is:

$$\frac{\partial \hat{y}_i}{\partial \ell_{i,c}} = -\pi_{i,c} \, (a_c - \hat{y}_i) \tag{34}$$

Reducing the NLL of an anchor above the current prediction increases $\hat{y}_i$, whereas reducing the NLL of an anchor below the current prediction decreases it. The mapping is therefore fully differentiable while maintaining a prediction direction consistent with the numerical ordering of the anchors.

Stability of soft expectation. Let $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \Delta^{C-1}$ denote two semantic distributions with predictions $\hat{y} = \mathbf{a}^\top \boldsymbol{\pi}$ and $\hat{y}' = \mathbf{a}^\top \boldsymbol{\pi}'$. Since $\sum_c (\pi_c - \pi'_c) = 0$, let $m = (a_1 + a_C)/2$. By Hölder's inequality:

$$\begin{aligned} |\hat{y} - \hat{y}'| &= \left|(\mathbf{a} - m\mathbf{1})^\top (\boldsymbol{\pi} - \boldsymbol{\pi}')\right| \\ &\leq \|\mathbf{a} - m\mathbf{1}\|_\infty \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1 \\ &\leq \frac{a_C - a_1}{2} \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1 \end{aligned} \tag{35}$$

Therefore, bounded perturbations of the semantic distribution induce bounded changes in the continuous prediction, demonstrating that the soft-expectation mapping is stable with respect to distributional variations.

A.13 Theoretical Analysis of Distance-Aware Supervision

To analyze the relationship between distance-aware supervision and continuous regression, we define the expectation of the soft target distribution $\boldsymbol{\tau}_i$ over the semantic anchors as:

$$\mu_i^\tau = \sum_{c=1}^{C} \tau_{i,c} a_c \tag{36}$$

From Eq. (13), $\hat{y}_i$ is the expectation of the predicted distribution $\boldsymbol{\pi}_i$ over the same anchor set. For two probability distributions P and Q on the real line, the Kantorovich–Rubinstein dual formulation of the Wasserstein-1 distance is:

$$W_1(P, Q) = \sup_{\mathrm{Lip}(f) \leq 1} |\mathbb{E}_P[f(A)] - \mathbb{E}_Q[f(A)]| \tag{37}$$

Since the identity function $f(a) = a$ is 1-Lipschitz, setting $P = \boldsymbol{\pi}_i$ and $Q = \boldsymbol{\tau}_i$ gives:

$$|\hat{y}_i - \mu_i^\tau| \leq W_1(\boldsymbol{\pi}_i, \boldsymbol{\tau}_i) \tag{38}$$

Therefore, minimizing the Wasserstein discrepancy directly constrains the deviation between the expectations of the predicted and target distributions. Combining this result with the triangle inequality yields:

$$|\hat{y}_i - y_i| \leq W_1(\boldsymbol{\pi}_i, \boldsymbol{\tau}_i) + |\mu_i^\tau - y_i| \tag{39}$$

The first term measures the discrepancy between the predicted and target distributions while accounting for distances between semantic anchors. The second term represents the approximation bias between the soft-target expectation and the original continuous label. Since the soft target is defined over a finite anchor set, μ_i^τ is not necessarily equal to y_i. This bias may be more pronounced near the boundaries of the anchor range, where the finite support truncates the soft target distribution and shifts its expectation toward the interior.

The Wasserstein and MAE objectives are therefore complementary. The Wasserstein loss aligns the predicted and target distributions while accounting for distances between anchors, whereas MAE directly supervises the prediction toward the exact continuous target. Their combination jointly promotes accurate regression and distributional consistency over the continuous sentiment space.