Decision-Level Ordinal Modeling for Multimodal Essay Scoring with Large Language Models

Anonymous ACL submission

Abstract

Automated essay scoring (AES) predicts multiple rubric-defined trait scores for each essay, where each trait follows an ordered discrete rating scale. Most LLM-based AES methods cast scoring as autoregressive token generation and obtain the final score via decoding and parsing, making the decision implicit. This formulation is particularly sensitive in multimodal AES, where the usefulness of visual inputs varies across essays and traits. To address these limitations, we propose **D**ecision-**L**evel **O**rdinal **M**odeling (**DLOM**), which makes scoring an explicit ordinal decision by reusing the LM head to extract score-wise logits on predefined score tokens, enabling direct optimization and analysis in the score space. For multimodal AES, **DLOM-GF** introduces a gated fusion module that adaptively combines textual and multimodal score logits. For text-only AES, **DLOM-DA** adds a distance-aware regularization term to better reflect ordinal distances. Experiments on the multimodal EssayJudge dataset show that DLOM improves over a generation-based SFT baseline across scoring traits, and DLOM-GF yields further gains when modality relevance is heterogeneous. On the text-only ASAP/ASAP++ benchmarks, DLOM remains effective without visual inputs, and DLOM-DA further improves performance and outperforms strong representative baselines.

1 Introduction

Automated Essay Scoring (AES) is a long-standing problem in natural language processing (Williamson et al., 1999; Smolentzov, 2013), aiming to automatically assess the quality of student essays according to predefined scoring rubrics (Misgna et al., 2024), often across multiple scoring traits rather than a single holistic score (Ridley et al., 2021; Kumar et al., 2022; Do et al., 2024a). Recent advances in large language models (LLMs) have significantly improved semantic understanding and instruction following, making them a natural choice for AES. As a result, a growing body of work has explored LLM-based essay scoring (Ridley et al., 2021; Ramesh and Sanampudi, 2022; Xia et al., 2024; Chu et al., 2025), including extensions to multimodal settings where essays are accompanied by visual inputs such as charts or diagrams, which further complicates the scoring decision (Su et al., 2025a,b).

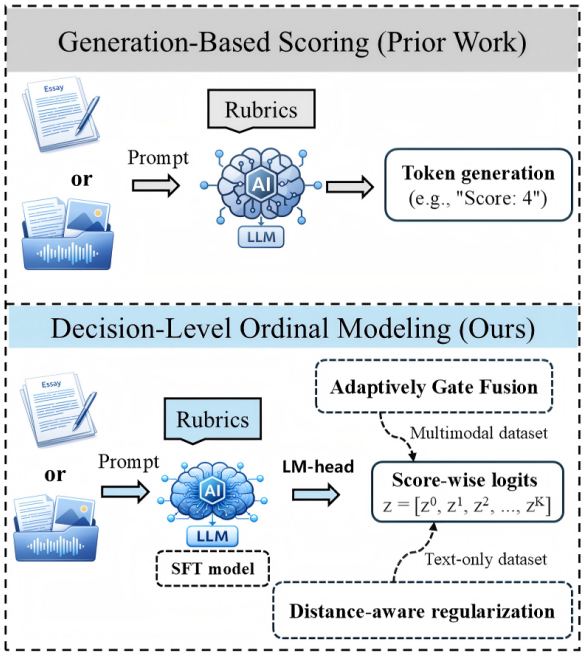

Figure 1: Comparison between generation-based scoring and decision-level ordinal modeling (DLOM).

Most existing LLM-based AES approaches formulate scoring as a generation problem (Do et al., 2024a,b), where the model is prompted to produce a score token or short textual response corresponding to a rubric-defined grade. This formulation is appealing due to its simplicity and compatibility with instruction-tuned LLMs (Pack et al., 2024; Song et al., 2024; Bui and Barrot, 2025). However, generation-based scoring makes the scoring

decision implicit, as the final score emerges as a byproduct of language generation (Liu et al., 2023). Consequently, the decision process is difficult to control, analyze, or adapt, and is often sensitive to prompt design, decoding strategies, and tokenization artifacts—factors that are largely orthogonal to the scoring task itself (Wang et al., 2018).

In particular, AES is inherently ordinal: score categories follow a fixed and ordered structure, and errors between adjacent score levels are typically less severe than those between distant levels (Johnson, 1996; Taghipour and Ng, 2016; Alikaniotis et al., 2016). Motivated by this, we adopt a decision-based formulation of essay scoring to better match the ordinal nature of rubric-based assessment. We further argue that, especially with LLMs that provide strong semantic representations, separating semantic understanding from decision making offers a more appropriate and general modeling perspective for essay scoring.

In this work, we propose Decision-Level Ordinal Modeling (DLOM) for LLM-based essay scoring. Figure 1 illustrates the difference between generation-based scoring and our proposed decision-level ordinal modeling (DLOM). Building on a supervised fine-tuned (SFT) LLM, we cast score prediction as an explicit decision over a predefined ordered label set. The model extracts score-wise logits over the ordinal categories, enabling direct optimization and analysis at the decision stage while preserving the semantic representations learned during SFT. We further extend DLOM to multimodal scoring with a decision-level gated fusion module (**DLOM-GF**) that adaptively combines text-only and multimodal score logits. For text-only benchmarks, we introduce a distance-aware regularization variant (**DLOM-DA**) to better respect ordinal distances between score levels.

Our main contributions are threefold:

- **Decision-level ordinal modeling for LLM-based AES.** We cast scoring as an explicit ordinal decision by predicting score-wise logits over predefined score levels, enabling direct optimization and analysis in the score space.
- **DLOM extensions for multimodal and text-only scoring.** We introduce DLOM-GF for decision-level multimodal gated fusion and DLOM-DA for distance-aware regularization in text-only settings.
- **Comprehensive evaluation on multimodal and text-only benchmarks.** We demonstrate the effectiveness and generality of the proposed formulation on both multimodal and text-only essay scoring benchmarks, achieving consistent improvements over generation-based SFT baselines.

2 Related Work

2.1 Automated Essay Scoring

Automated essay scoring (AES) has progressed from feature-engineered systems based on surface, syntactic, and discourse cues (Page, 1966; Dikli, 2006; Dong and Zhang, 2016; Hou et al., 2025) to neural representation learning with CNN/RNN architectures and pre-trained language models (Dong et al., 2017; Shibata and Uto, 2025). Recent transformer-based models further improve scoring by leveraging contextual representations and transfer learning (Devlin et al., 2019; Yang et al., 2020; Azhari et al., 2024), enabling end-to-end learning of essay representations and scoring functions.

2.2 LLM-Based Essay Scoring

With the rise of large language models (LLMs), many recent AES methods use instruction-following LLMs as backbones (Dong et al., 2021; Li and Liu, 2024). Most work adopts a generation-based interface, prompting the model to output a score token or short response aligned with rubric-defined grades (Zhou et al., 2023; Li and Pan, 2025). This approach is simple and often requires no architectural changes, but the final label is typically obtained by decoding and parsing generated text, making it sensitive to prompting, decoding, and tokenization details. Several studies improve robustness via prompt design, calibration, or decoding constraints (Wiher et al., 2022; Liu et al., 2024), while largely retaining the generation paradigm.

2.3 Ordinal Modeling for Scoring Tasks

Ordinal modeling has been recognized as important for scoring and assessment tasks, where score categories follow a fixed order and misclassification costs are asymmetric (Frank and Hall, 2001; Gutiérrez et al., 2016). Prior work outside the LLM context has explored ordinal regression, ranking-based losses, and distance-aware objectives to better reflect the structure of scoring scales (McCullagh, 1980; Herbrich et al., 2000; Suárez et al., 2021). In LLM-based AES, however, ordinal structure is often only implicitly encoded through tex-

Paradigm	Decision interface	Multimodal fusion
Generation-based LLM AES	Decode and parse score text	Prompt/decoding dependent
Pooling-head classifier	Pool(H) + $(K+1)$-way classification head	Usually representation-level
DLOM (ours)	**LM-head score-logit extraction on fixed score tokens**	**Decision-level gated fusion**

Table 1: Positioning of our approach against common decision interfaces used in AES.

tual descriptions of score levels or prompt instructions (Stahl et al., 2024). Such an explicit decision interface over ordered labels has received relatively limited attention in LLM-based AES, motivating our decision-level ordinal formulation.

Positioning. Table 1 summarizes common decision interfaces in LLM-based AES. Compared to generation-and-parse scoring and pooling-head classifiers, DLOM reuses the LM head to extract score-wise logits on fixed score tokens and makes predictions directly in the ordered score space, avoiding both decoding-time artifacts and an additional classification head.

2.4 Multimodal Essay Scoring

Beyond text-only AES, recent work has explored multimodal essay scoring scenarios in which essays are accompanied by visual inputs such as charts, tables, or diagrams (Su et al., 2025a). Most approaches perform multimodal fusion at the representation level, combining textual and visual information via concatenation, attention mechanisms, or cross-modal transformers (Su et al., 2025b). Since modality relevance can vary across essays and traits, representation-level fusion may be sensitive to heterogeneous modality contributions. Decision-level fusion that operates on score predictions has been less explored in AES, and offers an alternative interface for controlling modality influence and improving interpretability.

3 Methodology

3.1 Problem Formulation

Given an essay e and an optional visual input v, the goal of automated essay scoring is to predict a discrete score y from a fixed, ordered set of score levels $\mathcal{Y} = \{0, 1, \ldots, K\}$. The score levels follow an inherent ordinal structure, where misclassifications between adjacent levels are less severe than those between distant levels. We assume access to a supervised fine-tuned (SFT) large language model that encodes the semantic content of the essay. Our objective is to design a decision mechanism that operates on top of this model and explicitly accounts for the ordinal nature of the scoring task at the decision stage.

3.2 Decision-Level Ordinal Modeling

Instead of formulating essay scoring as a language generation task, we cast score prediction as an explicit decision over a fixed set of ordered score levels. A standard AES classifier typically adds a $(K+1)$-way linear head on top of pooled hidden states (e.g., [CLS]) and predicts scores in that head space. In contrast, our decision-level formulation reuses the LLM's LM head (vocabulary projection) and derives score-wise logits directly from the model outputs without decoding. Concretely, we restrict the final-step vocabulary logits to a predefined set of score tokens, obtaining $\mathbf{z} \in \mathbb{R}^{K+1}$ as the decision space. This keeps the backbone unchanged while making the scoring decision explicit in an ordered score space, and also supports decision-level fusion for multimodal scoring.

Figure 2 provides an overview of the proposed framework. Given an input essay (and optional visual input), the SFT LLM produces final-step vocabulary logits via its LM head. We select the entries corresponding to the predefined score tokens to form the score-wise logit vector $\mathbf{z} = [z_0, z_1, \ldots, z_K] \in \mathbb{R}^{K+1}$. Formally, $\mathbf{z} = \mathrm{Logits}(e, v)_L[\mathcal{S}]$, where $\mathcal{S}$ denotes the score-token set and L is the final position. This formulation decouples semantic understanding from decision making: the LLM encodes essay semantics, while the scoring decision is made explicitly in the ordinal score space. At inference time, the predicted score is obtained by $\hat{y} = \arg\max_{k \in \mathcal{Y}} z_k$. A detailed comparison between generation-based scoring, CLS-head classification, and our decision-level interface is provided in Appendix A.

3.3 DLOM-GF for Multimodal AES

For multimodal essay scoring settings where both textual and visual inputs are available, the reliability of visual information varies across essays and scoring traits. We introduce a decision-level gating mechanism that adaptively combines textual

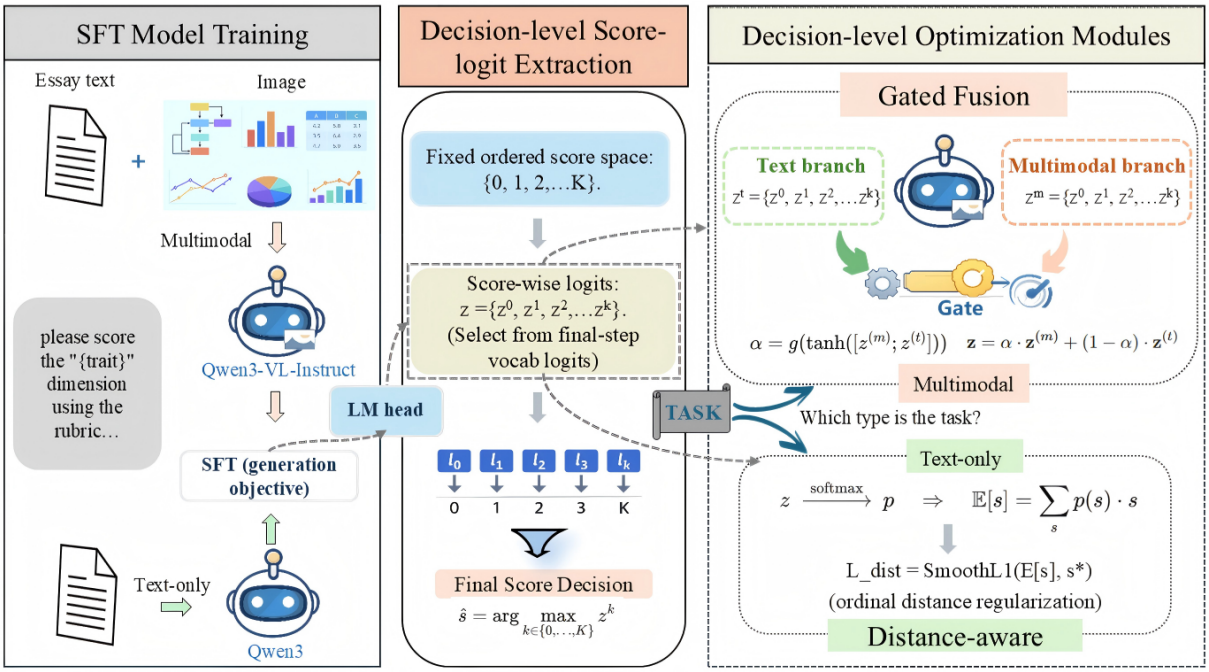

Figure 2: Overview of the proposed decision-level ordinal modeling framework. The framework consists of three stages: (i) supervised fine-tuning (SFT) for semantic encoding, (ii) decision-level score-logit extraction over an ordered score-token set, and (iii) task-specific decision-level objectives: decision-level gated fusion for multimodal scoring and distance-aware regularization for text-only scoring.

and multimodal predictions based on their score-level evidence. Specifically, we obtain two score-wise logit vectors: $\mathbf{z}^{(t)} \in \mathbb{R}^{K+1}$ from the text-only branch and $\mathbf{z}^{(m)} \in \mathbb{R}^{K+1}$ from the multimodal branch, where each logit corresponds to a candidate score level. We concatenate the two logit vectors and apply a non-linear transformation to produce a gating input:

$$\mathbf{h} = \tanh([\mathbf{z}^{(m)}; \mathbf{z}^{(t)}]).$$

A lightweight gating network $g(\cdot)$ then maps $\mathbf{h}$ to a scalar fusion weight:

$$\alpha = g(\mathbf{h}) \in [0, 1],$$

where $g(\cdot)$ is implemented as a linear layer followed by a sigmoid activation. The final score logits are computed as a convex combination in the ordinal score space:

$$\mathbf{z} = \alpha \cdot \mathbf{z}^{(m)} + (1 - \alpha) \cdot \mathbf{z}^{(t)}.$$

By conditioning the gate on score-wise logits rather than intermediate representations, the proposed mechanism operates entirely at the decision level. The gating function is learned implicitly through the scoring objective, allowing the model to estimate modality reliability in an instance-adaptive and task-aligned manner.

3.4 Training Objective

The model is trained to predict an ordinal score label from a fixed and ordered score set. Given the score-wise logits $\mathbf{z} \in \mathbb{R}^{K+1}$ produced by the decision-level score-logit extraction (or the gated fusion module in multimodal settings), we compute a categorical distribution $p = \mathrm{softmax}(\mathbf{z})$ and optimize the cross-entropy loss with respect to the gold score y:

$$\mathcal{L}_{\mathrm{CE}} = -\log p_y.$$

For multimodal essay scoring, the score logits $\mathbf{z}$ are obtained via the proposed decision-level gated fusion module, and the entire model is trained end-to-end by backpropagating $\mathcal{L}_{\mathrm{CE}}$ through both modality branches and the gating function.

In text-only settings, where decision-level fusion is not applicable, we additionally introduce a distance-aware regularization term to better capture the ordinal structure of the scoring scale. Specifically, we compute the expected score under the predicted distribution, $\mathbb{E}[s] = \sum_{k \in \mathcal{Y}} p_k \cdot k$, and penalize its deviation from the gold score using a SmoothL1 loss:

$$\mathcal{L}_{\mathrm{dist}} = \mathrm{SmoothL1}(\mathbb{E}[s], y).$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\, \mathcal{L}_{\text{dist}} + \beta(\lambda - 0.5)^2,$$

where $\lambda = \sigma(\lambda_{\text{logit}}) \in (0, 1)$ is a learnable weight and β is a small regularization coefficient. This auxiliary objective encourages ordinal consistency by assigning larger penalties to predictions that are farther away from the ground-truth score.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our approach on both multimodal and text-only multi-trait essay scoring benchmarks and treat each trait as an independent ordinal prediction task and report trait-wise results as well as their macro-averaged performance.

Multimodal Dataset. For multimodal essay scoring, we use the EssayJudge dataset (Su et al., 2025a), which consists of student essays accompanied by visual inputs such as charts or diagrams. Each essay is annotated with multiple trait-level scores according to a predefined rubric, covering diverse aspects of writing quality. This dataset is specifically designed to assess multimodal reasoning in essay scoring, where visual information may contribute unevenly across essays and traits (more information is in the Appendix B.1).

Text-only Datasets. For text-only evaluation, we use the combined ASAP/ASAP++ benchmark, which is a standard dataset for multi-trait automated essay scoring. The original ASAP dataset provides holistic scores for all prompts, while only a subset of prompts includes trait-level annotations. ASAP++ extends ASAP by providing trait scores for the remaining prompts (Mathias and Bhattacharyya, 2018, 2020), enabling comprehensive multi-trait evaluation across all prompts. The final dataset consists of eight writing prompts, each associated with its own scoring rubric and traits to be evaluated (more information is in the Appendix B.2). Following prior work, we use the combined ASAP/ASAP++ dataset for trait-level and cross-prompt evaluation.

Evaluation Metrics. We adopt Quadratic Weighted Kappa (QWK) as the primary evaluation metric (Cohen, 1968), which is widely used in AES to measure agreement between model predictions and human raters while accounting for the ordinal nature of score scales. More information is shown in Appendix B.3.

4.2 Experimental Setup

Backbone Models. We use Qwen3-VL-8B-Instruct (Qwen Team, 2025b) as the backbone language model for multimodal experiments and Qwen3-8B (Qwen Team, 2025a) for text-only experiments. Both models are initialized from publicly released checkpoints and further adapted to the essay scoring task through supervised fine-tuning. For further details on the model checkpoints, please refer to Appendix B.4.

Training Protocol. For both multimodal and text-only settings, we first perform generation-based supervised fine-tuning using a unified prompt template based on the official rubric descriptions. We then apply the proposed DLOM training on top of the SFT models, details of the prompt template and rubrics are provided in Appendix B.5.

In multimodal experiments, the text-only and multimodal branches are trained jointly, with score predictions from both branches combined through the decision-level gated fusion mechanism. The entire model, including the gating function, is optimized end-to-end under the same ordinal classification objective. All experiments are conducted using five-fold cross-validation, and results are reported by averaging performance across folds. For the text-only datasets (ASAP/ASAP++), we follow the previously established partitioning scheme (Taghipour and Ng, 2016). For the multimodal dataset, which lacks an existing standard split, we construct five folds at the essay level, ensuring that no essay appears in more than one fold. We do not apply early stopping during training. Instead, all models are trained for a fixed number of epochs under the same training configuration.

Inference. At inference time, for both text-only and multimodal settings, scores are predicted directly from score-wise logits, and each trait is evaluated independently.

4.3 Main Results on EssayJudge

To enable supervised training and reproducible comparison on EssayJudge, we establish an explicit data partition and report training-based results. Prior work (CAFES) evaluates EssayJudge mainly through inference-time agentic pipelines over off-the-shelf MLLMs (Su et al., 2025b). We therefore report their best-reported results in our table as a reference. Table 2 shows the main results on the multimodal EssayJudge dataset under the

Model	Lexical Level		Sentence Level				Discourse Level				Avg
	LA	LD	CH	GA	GD	PA	AC	JP	OS	EL	
CAFES	0.510	0.500	0.520	0.570	0.540	0.490	0.370	0.440	0.480	0.280	0.470
SFT-Gen (baseline)	0.595	0.547	**0.570**	0.569	0.591	0.503	0.205	0.404	0.486	0.455	0.492
DLOM	0.594	0.544	0.562	0.580	0.587	0.498	**0.251**	0.425	0.517	0.477	0.504
DLOM-GF	**0.624**	**0.551**	0.569	**0.589**	**0.613**	**0.506**	**0.251**	**0.447**	**0.527**	**0.479**	**0.516**

Table 2: Main results on the EssayJudge dataset under the multi-trait scoring setting. LA: lexical accuracy, LD: lexical diversity, CH: coherence, GA: grammatical accuracy, GD: grammatical diversity, PA: punctuation accuracy, AC: argument clarity, JP: justifying persuasiveness, OS: organizational structure, EL: essay length.

Model	Traits										Avg
	Content	PA	Lang	Nar	Org	Conv	WC	SP	Style	Voice	
HISK	0.679	0.697	0.605	0.659	0.610	0.527	0.579	0.553	0.609	0.489	0.601
MTL-BiLSTM	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.626
ArTS	0.730	0.751	0.698	0.725	0.672	0.668	0.679	0.678	0.721	0.570	0.689
SFT-Gen (baseline)	0.705	0.752	0.698	0.732	0.638	0.582	0.632	0.614	0.521	0.585	0.646
DLOM	0.738	0.776	0.716	0.753	0.692	0.655	0.659	0.662	0.615	0.581	0.685
DLOM-DA	**0.745**	**0.778**	**0.720**	**0.756**	**0.699**	**0.659**	**0.673**	**0.669**	**0.644**	**0.634**	**0.697**

Table 3: Trait-level results on the ASAP/ASAP++ benchmark. PA: prompt adherence, Lang: language, Nar: narrativity, Org: organization, Conv: conventions, WC: word choice, SP: sentence fluency.

multi-trait scoring setting. We compare three scoring paradigms: generation-based supervised fine-tuning (SFT-Gen, baseline), decision-level ordinal modeling (DLOM), and decision-level modeling with gated multimodal fusion (DLOM-GF). Compared to SFT-Gen, decision-level ordinal modeling results in a higher average QWK, while exhibiting differentiated effects across scoring traits. In particular, the decision-level formulation yields clear gains on discourse-level dimensions such as argument clarity and organizational structure, which require holistic reasoning over essay content. In contrast, improvements on lexical-level and sentence-level traits are comparatively limited, suggesting that the benefits of explicit ordinal decision modeling are more pronounced for higher-level scoring criteria.

Building upon the decision-level formulation, incorporating a gated multimodal fusion mechanism further improves overall performance and achieves the best average QWK among all compared methods. Beyond improving individual lower-level traits, the gating mechanism consistently strengthens performance on discourse-level dimensions, further stabilizing gains on traits that rely on holistic essay evaluation. These results suggest that adaptively integrating textual and visual evidence at the decision stage not only alleviates weaknesses in specific traits, but also reinforces the advantages of decision-level modeling for high-level essay scoring. You can find more information about main results in Appendix B.6

4.4 Text-only Results on ASAP/ASAP++

To examine whether the proposed decision-level formulation generalizes beyond the multimodal setting, we further evaluate our approach on the text-only ASAP/ASAP++ benchmark. For completeness, we also include several representative baselines (HISK (Cozma et al., 2018), MTL-BiLSTM (Kumar et al., 2022), and ArTS (Do et al., 2024a)). Overall, our variant DLOM-DA yields the best average performance among all compared methods.

Table 3 reports trait-level results on the dataset. Compared to the baseline, our DLOM leads to improvements on most traits and increases the average QWK from 0.646 to 0.685, indicating the effectiveness of explicitly modeling ordinal score decisions. Incorporating distance-aware regularization further improves performance and achieves the best average results among all compared methods, raising the average QWK to 0.697. Notably, the distance-aware variant yields consistent gains across all traits, suggesting that the additional ordinal regularization helps stabilize score predictions in the text-only setting.

The prompt-level results on the benchmark are in Table 4. The DLOM increases the average QWK from 0.676 to 0.713, and incorporating distance-

Model	Prompts								Avg
	1	2	3	4	5	6	7	8	
HISK	0.674	0.586	0.651	0.681	0.693	0.709	0.641	0.516	0.644
MTL-BiLSTM	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665
ArTS	0.708	0.706	0.704	0.767	0.723	0.776	0.749	0.603	0.717
SFT-Gen (baseline)	0.675	0.651	0.721	0.770	0.706	0.753	0.556	0.572	0.676
DLOM	**0.682**	**0.702**	0.735	0.789	**0.730**	0.778	0.664	0.621	0.713
DLOM-DA	0.681	0.696	**0.743**	**0.795**	0.728	**0.780**	**0.688**	**0.646**	**0.720**

Table 4: Prompt-level results on the ASAP/ASAP++ benchmark.

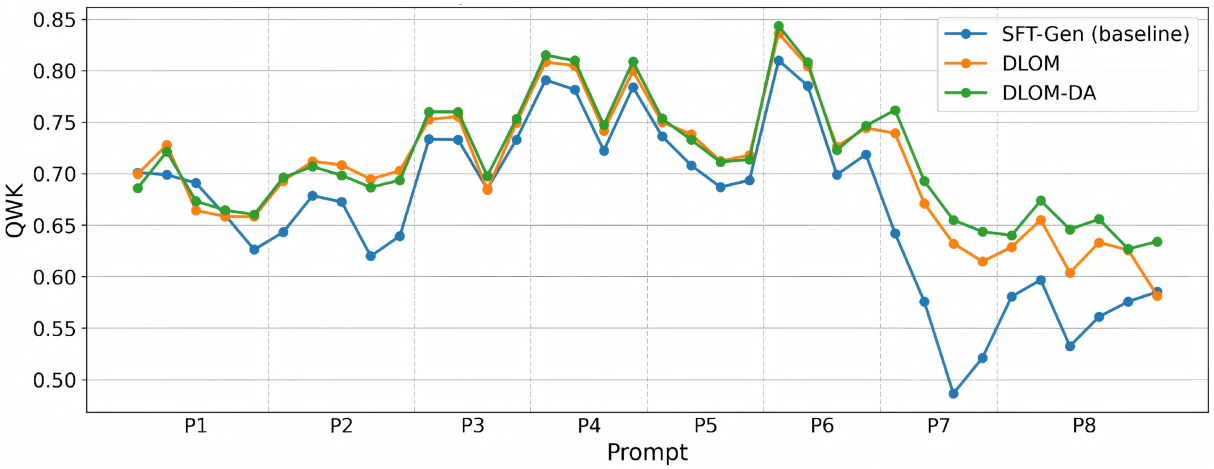

Figure 3: Prompt-wise QWK trends across different models on ASAP/ASAP++. Each point corresponds to a specific trait under a given prompt. Vertical dashed lines indicate boundaries between prompts.

aware regularization further improves performance to 0.720. Specifically, the decision-level ordinal modeling achieves gains on all prompts. Overall, the text-only results corroborate our findings in the multimodal setting. Even without visual inputs, explicitly modeling essay scoring at the decision level leads to consistent performance improvements, demonstrating that the proposed formulation's generalization ability. More details are provided in Appendix B.7.

4.5 Analysis and Discussion

Prompt-wise Analysis. To better understand the robustness of different modeling paradigms with respect to prompt variations, we conduct a prompt-wise analysis on the text-only dataset. Figure 3 presents QWK trends across prompts, where we evaluate the corresponding traits under each prompt. While absolute performance varies across prompts and traits, SFT-Gen exhibits noticeable fluctuations, with a sustained performance degradation on P7 and P8 (i.e., consistently lower QWK across the associated traits). In contrast, DLOM yields more consistent trends across prompts, substantially mitigating extreme degradations observed in the SFT baseline. This effect is further strengthened by incorporating distance-aware ordinal modeling, while DLOM-DA consistently maintains higher QWK scores across most prompts. These observations suggest that explicitly modeling scoring decisions in the ordinal space reduces sensitivity to prompt-specific variations and leads to more robust essay scoring behavior.

Method	Avg. QWK
SFT-Gen (baseline)	0.492
SFT + Ordinal Loss (w=0.01)	0.472
SFT + Ordinal Loss (w=0.1)	0.481
DLOM (ours)	**0.504**

Table 5: Effect of decision-level formulation on the multimodal EssayJudge dataset.

Necessity of Decision-level Formulation. A natural question is whether ordinal structure can be incorporated into generation-based essay scoring by simply adding an ordinal loss term, without modifying the underlying modeling formulation. In this setting, the ordinal loss is introduced as an auxiliary regularization term, with w controlling its relative weight to the token-level generation objective. To examine this alternative, we compare decision-level

Inference Strategy	LA	LD	CH	GA	GD	PA	AC	JP	OS	EL	Avg.
Text-only Decision	0.524	0.551	0.474	0.541	0.559	0.414	0.249	0.342	0.525	0.363	0.454
Multimodal-only Decision	0.616	0.520	0.544	0.580	0.602	0.499	0.242	0.416	0.515	0.472	0.501
DLOM-GF (ours)	**0.624**	**0.551**	**0.569**	**0.589**	**0.613**	**0.506**	**0.251**	**0.447**	**0.527**	**0.479**	**0.516**

Table 6: Per-trait comparison of different decision-level inference strategies on the multimodal EssayJudge dataset.

ordinal modeling with generation-based SFT augmented with ordinal loss under different values of w on the multimodal dataset. As shown in Table 5, across the tested loss weights, generation-based models with ordinal regularization consistently underperform decision-level ordinal modeling, and even fall below the original SFT baseline.

These results demonstrate that explicitly reformulating essay scoring as a decision over a fixed, ordered score space leads to a higher average QWK. Unlike loss-level ordinal constraints, which act as auxiliary signals on token-level generation, decision-level ordinal modeling embeds ordinality directly into the prediction space and the decision process itself. This formulation encourages the model to perform score-wise comparisons and reason explicitly over relative score preferences, rather than relying on indirect token-level supervision. These results suggest that, under our experimental setting, ordinal structure in essay scoring is more effectively leveraged when modeled at the decision level rather than imposed solely through loss-level regularization.

Effect of Decision-Level Fusion. Building on the decision-level formulation, we further investigate whether fusing modality-specific score logits can yield additional gains in multimodal scoring. Table 6 presents a per-trait comparison of different decision-level inference strategies on the EssayJudge dataset. Overall, decision-level fusion achieves the highest average QWK (0.516), outperforming both text-only (0.454) and multimodal-only (0.501) inference. This indicates that, even when the backbone and the decision objective are fixed, fusing modality-specific score logits at the decision stage provides a clear additional benefit.

At the trait level, decision-level fusion attains the best performance across all scoring dimensions. While multimodal-only inference generally outperforms text-only inference, it still consistently lags behind decision-level fusion. These results suggest that textual and visual cues provide complementary information, and that the adaptive decision-level gating mechanism can selectively leverage the more informative modality, leading to more reliable multimodal essay scoring decisions.

Design Implications. Our analysis suggests several practical insights for LLM-based essay scoring. First, in the text-only setting, decision-level ordinal modeling exhibits improved robustness to prompt variations, and additional distance-aware regularization can further enhance performance stability. Second, the key to leveraging ordinality lies in the decision-level formulation: by casting scoring as an explicit decision over a fixed ordered score set, ordinality is encoded directly in the prediction space and the decision mechanism. Third, for multimodal essay scoring, decision-level fusion with an adaptive gate provides an effective way to integrate textual and visual cues in the ordinal score space without requiring representation-level fusion. Overall, these observations highlight the importance of aligning modeling choices with the structure of the scoring decision itself.

5 Conclusion

In this work, we revisited LLM-based automated essay scoring from a decision-centric perspective. We reformulated scoring as an explicit ordinal decision over a fixed, ordered score set, enabling direct optimization in the score space without relying on decode-and-parse generation. On top of this formulation, we introduced a decision-level gated fusion module for multimodal AES and a distance-aware regularization variant for text-only scoring.

Empirically, decision-level ordinal modeling improves over generation-based SFT on the multimodal EssayJudge dataset, and gated fusion yields further gains when visual relevance is heterogeneous. On the text-only ASAP/ASAP++ benchmark, the same formulation remains effective, with distance-aware regularization providing additional improvements. Overall, our results highlight that explicitly modeling the ordinal nature of scoring can yield more reliable LLM-based assessment, and the same view may extend to other ordinal or decision-centric evaluation tasks.

Limitations

This work focuses on decision-level modeling for essay scoring and evaluates its effectiveness on a limited set of backbone models and datasets. Although the proposed formulation and fusion strategy demonstrate consistent improvements, our experiments are conducted primarily with Qwen-based large language models. Further validation on a broader range of architectures may be necessary to assess the generality of the approach.

In addition, while our decision-level fusion framework is applicable to multimodal settings, distance-aware regularization is only explored in the text-only scenario. We leave a systematic investigation of integrating additional ordinal constraints with multimodal fusion to future work.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Azhari Azhari, Agus Santoso, Anak Agung Putri Ratna, and Jasson Prestiliano. 2024. Optimization of aes using bert and bilstm for grading the online exams. *International Journal of Intelligent Engineering & Systems*, 17(5).

Ngoc My Bui and Jessie S Barrot. 2025. Chatgpt as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*, 30(2):2041–2058.

SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5796–5814.

J. Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia. Association for Computational Linguistics.

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

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Heejin Do, Yunsu Kim, and Gary Lee. 2024a. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666.

Heejin Do, Sangwon Ryu, and Gary Lee. 2024b. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1072–1077.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pages 153–162.

Lulu Dong, Lin Li, HongChao Ma, YeLing Liang, and 1 others. 2021. Automated chinese essay scoring using pre-trained language models. In *CS & IT Conference Proceedings*, volume 11. CS & IT Conference Proceedings.

Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *Machine Learning: ECML 2001*, pages 145–156, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pedro Antonio Gutiérrez, María Pérez-Ortiz, Javier Sánchez-Monedero, Francisco Fernández-Navarro, and César Hervás-Martínez. 2016. Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 2000. Large margin rank boundaries for ordinal regression. In *Advances in Large-Margin Classifiers*. The MIT Press.

Zhaoyi Joey Hou, Alejandro Ciuba, and Xiang Lorraine Li. 2025. Improve llm-based automatic essay scoring with linguistic features. *arXiv preprint arXiv:2502.09497*.

Valen E Johnson. 1996. On bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association*, 91(433):42–51.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495.

Wenchao Li and Haitao Liu. 2024. Applying large language models for automated essay scoring for non-native japanese. *Humanities and Social Sciences Communications*, 11(1):1–15.

Xia Li and Wenjing Pan. 2025. Ceaes: Bidirectional reinforcement learning optimization for consistent and explainable essay assessment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26267–26279.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating llm-based evaluator. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pages 2638–2656.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91.

Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Haile Misgna, Byung-Won On, Ingyu Lee, and Gyu Sang Choi. 2024. A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58(2):36.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Qwen Team. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qwen Team. 2025b. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

Takumi Shibata and Masaki Uto. 2025. Cross-prompt automated essay scoring via reinforcement learning-based data valuation. *IEEE Access*.

André Smolentzov. 2013. Automated essay scoring: Scoring essays in swedish.

Yishen Song, Qianta Zhu, Huaibo Wang, and Qinhua Zheng. 2024. Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, 17:1880–1890.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Jiamin Su, Yibo Yan, Fangteng Fu, Zhang Han, Jingheng Ye, Xiang Liu, Jiahao Huo, Huiyu Zhou, and Xuming Hu. 2025a. EssayJudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6363–6389, Vienna, Austria. Association for Computational Linguistics.

Jiamin Su, Yibo Yan, Zhuoran Gao, Han Zhang, Xiang Liu, and Xuming Hu. 2025b. Cafes: A collaborative multi-agent framework for multi-granular multimodal essay scoring. *Preprint*, arXiv:2505.13965.

Juan Luis Suárez, Salvador García, and Francisco Herrera. 2021. Ordinal regression with explainable distance metric learning based on ordered sequences: Extended abstract. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–2.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuan-Jing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 791–797.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012.

David M Williamson, Isaac I Bejar, and Anne S Hone. 1999. 'mental model'comparison of automated and human scoring. *Journal of Educational Measurement*, 36(2):158–184.

Wei Xia, Shaoguang Mao, and Chanjing Zheng. 2024. Empirical study of large language models as automated essay scoring tools in english composition_taking toefl independent writing task for example. *arXiv preprint arXiv:2401.03401*.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, Xiaodong He, and 1 others. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Proceedings of ACL 2020*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Additional Clarification on Decision Interfaces

This appendix clarifies how our decision-level formulation differs from two common alternatives for LLM-based AES: generation-based scoring and standard classification heads.

Generation-based scoring. A generation-based scorer predicts a score by decoding a textual output (e.g., a score token or a short phrase) and then mapping the decoded text to an ordinal label. Formally, it relies on $\hat{y} = \mathrm{Parse}(\mathrm{Decode}(p_\theta(\cdot \mid e, v)))$, which introduces sensitivity to decoding strategies (e.g., sampling/greedy) and prompt/output formatting.

CLS-head classification (with/without ordinal losses). A standard classifier treats the LLM as an encoder, pools hidden states, and applies an additional $(K{+}1)$-way head:

$$\mathbf{o} = W_{\text{cls}} \cdot \mathrm{pool}(H) + b, \quad \hat{y} = \arg\max_k o_k,$$

optionally augmented with ordinal regression objectives (e.g., CORAL-style losses or distance-based regularizers) on top of $\mathbf{o}$. This design introduces a separate decision space parameterized by W_{cls}, which is not tied to the LM output distribution.

Ours: decision-level score-logit extraction. In contrast, we reuse the LM head and make the decision directly in a predefined ordered score space by restricting the final-step vocabulary logits to a fixed score-token set $\mathcal{S}$:

$$\mathbf{z} = \mathrm{Logits}(e, v)_L[\mathcal{S}] \in \mathbb{R}^{K+1}, \quad \hat{y} = \arg\max_{k \in \mathcal{Y}} z_k.$$

This removes the need for decoding and avoids introducing an additional classification head, while yielding an explicit score-wise decision representation. Operating in this score-logit space also enables our decision-level gated fusion, which adaptively combines modality-specific evidence via a convex combination of $\mathbf{z}^{(m)}$ and $\mathbf{z}^{(t)}$ under the same decision objective. Moreover, this interface is LLM-native: it leverages the model's learned token-preference distribution and avoids relying on prompt- or decoding-specific behaviors when producing a final score.

B Experiments

B.1 EssayJudge

Table 7 shows the detailed information about EssayJudge dataset (Su et al., 2025a). It consists of 1,054 multimodal essays written as part of IELTS Writing Task 1. In this task, candidates are required to write a report based on visual data such as charts, graphs, or diagrams. The dataset is specifically designed to assess multimodal reasoning in automated essay scoring, where visual inputs play a significant role in the essay's content and quality.

Statistic	**Number**
Total Multimodal Essays	1,054
Image Type	
- Single-Image	703 (66.7%)
- Multi-Image	351 (33.3%)
Multimodal Essay Type	
- Flow Chart	305 (28.9%)
- Bar Chart	211 (20.0%)
- Table	153 (14.5%)
- Line Chart	145 (13.8%)
- Pie Chart	71 (6.7%)
- Map	62 (5.9%)
- Composite Chart	107 (10.2%)

Table 7: Key statistics of EssayJudge dataset.

B.2 ASAP/ASAP++

We use the open-sourced `ASAP/ASAP++`, as summarized in Table 8, different prompts are assessed with distinct traits, each having varied score ranges (Mathias and Bhattacharyya, 2018, 2020). It consists of English essays written by American 7th to 10th-grade high school students across eight prompts.

Prompt	# Essays	Traits	Score Range
1	1785	Content, WC, Org, SF, Conv	1 - 6
2	1800	Content, WC, Org, SF, Conv	1 - 6
3	1726	Content, PA, Nar, Lang	0 - 3
4	1772	Content, PA, Nar, Lang	0 - 3
5	1805	Content, PA, Nar, Lang	0 - 4
6	1800	Content, PA, Nar, Lang	0 - 4
7	1569	Content, Org, Conv, Style	0 - 6
8	723	Content, WC, Org, SF, Conv, Voice	2 - 12

Table 8: Composition of the ASAP/ASAP++ combined dataset. The prompt defines the writing theme. Abbreviations: WC (Word Choice), Org (Organization), SF (Sentence Fluency), Conv (Conventions), PA (Prompt Adherence), Nar (Narrativity), Lang (Language).

Model	Source	URL
Qwen3-VL-8B-Instruct	local checkpoint	https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
Qwen3-8B	local checkpoint	https://huggingface.co/Qwen/Qwen3-8B

Table 9: Model sources for the multimodal and text-only essay scoring tasks.

B.3 Quadratic Weighted Kappa (QWK)

Quadratic Weighted Kappa (QWK) is a commonly used metric in automated essay scoring (AES) to measure the agreement between model predictions and human raters. It accounts for the ordinal nature of the score scales, where errors between adjacent score levels are less severe than those between distant levels.

The QWK is defined as:

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} \cdot O_{ij}}{\sum_{i,j} w_{ij} \cdot E_{ij}}$$

where: - O_{ij} is the observed agreement between categories i and j, - E_{ij} is the expected agreement between categories i and j, - w_{ij} is the weight assigned to the difference between score levels i and j.

The weight w_{ij} is typically defined as:

$$w_{ij} = \frac{(i-j)^2}{(K-1)^2}$$

where K is the number of possible score categories.

B.4 Model Selecting and Sources

For our experiments, we selected two large language models: Qwen3-VL-8B-Instruct and Qwen3-8B, chosen for their distinct strengths in multimodal and text-only tasks, respectively.

Qwen3-VL-8B-Instruct : This is a multimodal model designed to handle both textual and visual inputs, making it ideal for essay scoring tasks where essays are accompanied by images, charts, or other forms of visual data. And its open-source availability makes it accessible for replication and future work in this area.

Qwen3-8B : This is a text-only model, suitable for tasks that involve only textual input. We use it as a baseline to compare the performance of our multimodal model, allowing us to isolate the effect of incorporating visual information into the scoring process.

These two models represent a balanced comparison between multimodal and text-only settings, allowing us to evaluate the robustness and generality of our proposed method across different types of inputs. Table 9 shows their checkpoints' source.

B.5 Prompt Template and Rubrics

Training on multimodal dataset. The training was conducted using a unified prompt template as shown in Figure 4 and the rubric for each trait is publicly available and can be accessed in the original dataset repository (Su et al., 2025a).

```
    prompt_template = """
You are an IELTS examiner. Based on the image, question and essay, please
score the "{trait}" dimension using the rubric.


Rubric: {rubric}
Essay title: "{question}"
Essay: "{essay}"
Image: <image>

Your only output must be a single number (the score) from the rubric. No other
text.


Score:
    """
```

Figure 4: Prompt Template for Multimodal Dataset

Training on text-only datasets. We train all text-only models using a unified prompt template (Figure 5). Trait rubrics and prompt definitions for ASAP/ASAP++ are publicly available.[1]

```
prompt = (
        f"You must score the essay's \"{trait}\" dimension using the rubric.\n"
        f"Rubric: {rubric}\n"
        f"Essay: {essay}\n"
        f"Your only output must be a single number (the score) from the rubric.\n"
        f"Score:"
)
```

Figure 5: Prompt Template for Text-only Dataset

B.6 Main Results

We observe relatively large variance across cross-validation folds for several traits, including the generation-based baseline. This variance is mainly attributed to the limited size and heterogeneous nature of the EssayJudge dataset, rather than instability introduced by the proposed methods. Importantly, decision-level modeling and gated fusion do not exhibit higher variance than the baseline,

[1]ASAP: https://www.kaggle.com/c/asap-aes. ASAP++: https://lwsam.github.io/ASAP++/Irec2018.html.

Model	Content	PA	Lang	Nar	Org	Conv	WC	SP	Style	Voice	Avg
SFT (Generation)	±0.031	±0.026	±0.026	±0.025	±0.051	±0.066	±0.044	±0.049	±0.030	±0.056	±0.040
Decision-level (Ordinal)	±0.026	±0.025	±0.028	±0.025	±0.036	±0.036	±0.044	±0.032	±0.028	±0.083	±0.036
Decision-level + Distance-aware	±0.030	±0.023	±0.035	±0.031	±0.035	±0.038	±0.037	±0.032	±0.020	±0.083	±0.036

Table 10: Standard deviation of QWK across cross-validation folds on the ASAP/ASAP++ benchmark, reported at the trait level.

Model	1	2	3	4	5	6	7	8	Avg
SFT (Generation)	±0.053	±0.036	±0.032	±0.028	±0.021	±0.017	±0.054	±0.057	±0.037
Decision-level (Ordinal)	±0.020	±0.033	±0.033	±0.018	±0.035	±0.016	±0.034	±0.054	±0.031
Decision-level + Distance-aware	±0.029	±0.031	±0.040	±0.020	±0.036	±0.017	±0.030	±0.054	±0.032

Table 11: Standard deviation of QWK across cross-validation folds on the ASAP/ASAP++ benchmark, reported at the prompt level.

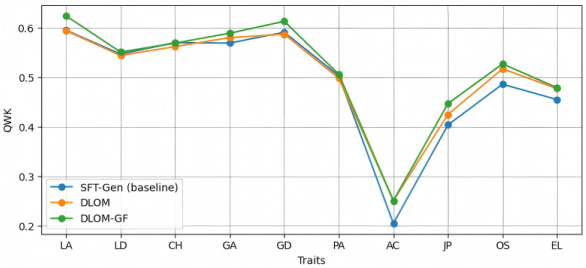

Figure 6: Trait-wise QWK comparison on the Essay-Judge dataset.

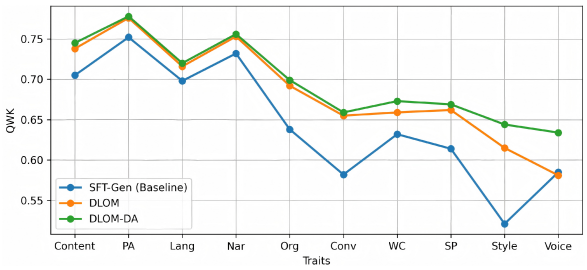

Figure 7: Trait-wise QWK trends on the ASAP/ASAP++ benchmark.

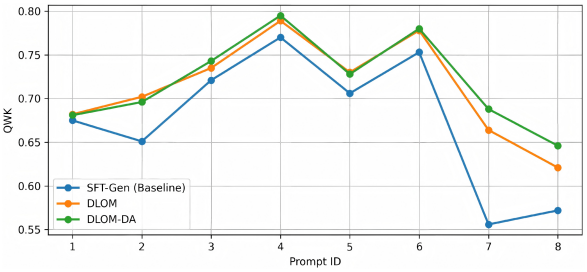

Figure 8: Prompt-wise QWK trends on the ASAP/ASAP++ benchmark.

indicating comparable stability under the same experimental setting.

Figure 6 provides a visualization of the trait-level QWK results reported in Table 2. The figure highlights consistent improvements of decision-level modeling over the generation-based baseline across most traits, with more pronounced gains on discourse-level dimensions. The gated fusion mechanism further strengthens performance, particularly on traits with lower baseline scores.

B.7 Text-only Results on ASAP/ASAP++

Variance analysis. Tables 10 and 11 report the standard deviation of QWK across cross-validation folds at the trait and prompt levels. Across both views, decision-level modeling and the distance-aware variant achieve higher mean performance without increasing variability: the average standard deviation remains comparable to the generation baseline. This suggests that the observed improvements are not obtained at the cost of reduced stability.

Visualizations of text-only results. Figures 7 and 8 visualize the trait-level and prompt-level QWK results from Tables 3 and 4, respectively. Overall, decision-level modeling improves over the generation-based baseline on most traits and prompts, and the distance-aware variant provides further (often more consistent) gains. Improvements are particularly pronounced on higher-level traits that require holistic judgments of content and organization.