APPENDIX

A. Ablation Study

To systematically evaluate the contribution of each component, we conducted a comprehensive ablation study, with results summarized in Table 1.

The analysis reveals that models trained on a single modality heavily rely on visual information. The image-only model established a strong baseline (Pearson's r = 0.952), significantly outperforming both the text-only (r = 0.899) and MiniLM-only (r = 0.889) counterparts. This underscores that visual features are the most critical component for assessing artistic creativity, though language-based cues alone provide meaningful, albeit insufficient, information.

Augmenting the visual stream with language features led to substantial performance gains. Fusing image data with either text or MiniLM embeddings boosted the Pearson correlation to 0.959. In contrast, the model combining only text and MiniLM, while better than either language model alone, lagged considerably behind (r = 0.901), further reinforcing the primacy of the visual modality.

Finally, the integration of all three modalities (image, text, and MiniLM) achieved the best overall performance across all metrics: the lowest MAE of 4.64, the highest Pearson's r of 0.965, and the top accuracy of 95.3%. This demonstrates a clear synergistic effect where textual and semantic embeddings effectively complement the primary visual information. These results validate our core hypothesis that a multimodal fusion approach is optimal, with visual data as the foundation and language-based features providing crucial enhancements for a more nuanced interpretation.

Table 1: Ablation study results for different modality combinations. ✓denotes the inclusion of a modality. Best results are in **bold**.

Input Modality			**Performance**		
Image	**Text**	**MiniLM**	**MAE (↓)**	**Pearson (r) (↑)**	**Accuracy (↑)**
✓			5.47	0.952	94.5%
	✓		8.13	0.899	91.9%
		✓	8.39	0.889	91.6%
✓	✓		5.05	0.959	94.9%
✓		✓	5.20	0.959	94.8%
	✓	✓	7.82	0.901	92.2%
✓	✓	✓	**4.64**	**0.965**	**95.3%**