IMPROVING MULTIMODAL MODELS IN AUTOMATED ARTWORK EVALUATION WITH OPTIMIZED SHOT SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated artwork evaluation with multimodal large language models (MLLMs) is useful for art education and analysis, but few-shot performance can be unreliable when the examples supplied in context are not representative. We study an optimized shot-selection strategy for improving MLLM-based artwork evaluation. Specifically, we compare random selection with K-Means and K-Medoids clustering, selecting representative artwork examples as few-shot demonstrations. Experiments on ratings of 337 student artworks show that K-Medoids generally provides more stable correlations with expert ratings and is especially effective for Color Richness and Picture Organization. The results also show that the number of shots has dimension-specific effects: additional examples can help some dimensions while harming others. These findings indicate that representative, outlier-resistant shot selection is an important component of reliable automated artwork evaluation.

1 INTRODUCTION

Multimodal large language models (MLLMs) have become effective general-purpose systems for image recognition, visual question answering, text-to-image generation, and cross-modal retrieval (Antol et al., 2015; Ramesh et al., 2021; Faghri et al., 2017). Recent work has also explored MLLMs for assisted artwork evaluation (Zheng et al., 2025). Models such as GPT-4 and Flamingo can integrate visual and linguistic information and perform complex tasks with natural-language interaction. Their broad capabilities have led to growing interest in using MLLMs for evaluation, but strong general performance does not imply reliable domain-specific judgment. In particular, MLLM outputs depend on the quality, ordering, and selection of the examples provided in context. This dependence introduces uncertainty in evaluation tasks, where a fluent answer can still be poorly grounded in domain knowledge.

Automated artwork evaluation is a valuable but challenging application. Schools often lack enough art teachers to provide detailed feedback on every student's work, while parents may not have the training required to evaluate early-stage artwork. An automated system could help provide timely feedback and support the development of artistic literacy. However, art evaluation is subjective and multidimensional: composition, color, line, technique, and transformation all contribute to a judgment. Meaning can also depend on historical context, cultural significance, and artistic style. Evaluators' aesthetic preferences and professional backgrounds further affect their ratings. These factors require an evaluation model to balance consistency with sensitivity to visual and cultural nuance. Existing evidence from essay evaluation suggests that MLLMs can still lag behind human experts in specialized assessment settings (Bui & Barrot, 2024).

Few-shot learning can improve adaptation to a new task with only a small number of annotated examples (Brown et al., 2020). It has been used for segmentation, generation, reasoning, and medical image analysis (Zhu et al., 2024; Bommasani et al., 2021; Kotia et al., 2021; Zhang et al., 2021a). Yet a random set of demonstrations is not guaranteed to improve an MLLM. In a long-tail setting, where many task configurations are rare, the model becomes particularly dependent on the available examples. A misleading, redundant, or out-of-distribution example can pull inference away from the intended rubric. The order of demonstrations can also change the inference path, so adding a second

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.

Table 1: Advantages of direct scoring in MLLMs.

Model	Advantage	Performance data
GPT-4 (Vision-enabled)	High scalability for large-scale tasks and efficient scoring	Processes 500 samples/min with consistent scoring reliability
Flamingo	Consistent scoring on structured tasks with reduced variability	Processes 400 samples/min; 78% alignment with expert evaluations
GPT-4	Reduced inter-rater variability using aggregated scoring rubrics	Variance reduced by 15–20% compared with baseline models
All MLLMs	Standardized scoring across domains with improved fairness and reproducibility	Suitable for high-throughput applications

example may make a prediction worse even when the first example helped. Prior example-selection research has focused mainly on text-only LLMs, including retrieval and uncertainty-ranking methods (Yu et al., 2024); these methods do not directly address the visual diversity and outliers found in artwork datasets.

This paper investigates whether clustering-based shot selection can improve MLLM artwork evaluation. We compare three strategies: random selection, K-Means selection, and K-Medoids selection. Each method chooses one, three, or five representative shots from a pool of artwork examples, which are then supplied to an MLLM. We compare the resulting ratings with expert ratings using Pearson correlation across nine artwork dimensions. Our main observations are that (i) shot selection has a strong dimension-dependent effect, (ii) K-Medoids is more stable than the other strategies in the presence of outliers, and (iii) representative examples can improve alignment without requiring many shots. The resulting analysis provides a practical basis for selecting demonstrations in automated artwork evaluation.

2 Related Work

2.1 Scoring mechanisms in MLLMs

MLLMs have been studied as direct scoring systems for art evaluation, essay grading, image quality assessment, and other structured tasks. Their main advantages are scalability, consistent application of a predefined rubric, and fast evaluation. The original study reports that GPT-4 (Vision-enabled) and Flamingo can process large numbers of samples while maintaining consistent scoring, and that aggregated scoring can reduce inter-rater variability (Achiam et al., 2023; Alayrac et al., 2022). Such systems may be useful when a large number of submissions must receive feedback under a common rubric.

Direct scoring also has important limitations. MLLMs may lack cultural, historical, and contextual understanding, which can produce incorrect interpretations of culturally loaded symbols. Visual ChatGPT is one example of a system studied for visual dialogue and editing (Wu et al., 2023). MLLMs can be inconsistent on ambiguous and creative tasks, and their outputs may inherit biases from training data. Multimodal inference is also computationally expensive, which limits deployment in resource-constrained settings. For these reasons, an MLLM score should be interpreted as an additional assessment signal rather than an unconditional replacement for expert judgment. Table 1 summarizes the advantages described in the supplied manuscript, while Table 2 summarizes the corresponding limitations.

2.2 Few-shot selection methods

Few-shot performance depends on selecting examples that are diverse, relevant, and informative. Active learning queries examples for which the model is uncertain, while uncertainty sampling focuses on low-confidence cases and underrepresented regions of the data distribution (Ren et al., 2018; Xiong et al., 2016). Clustering-based selection groups similar examples and chooses representatives, which can preserve diversity in multimodal settings (Snell et al., 2017). Task-specific embeddings can further align retrieval with the target task and reduce irrelevant context (Raffel et al., 2020).

Table 2: Limitations of MLLMs for nuanced evaluation.

Model	Limitation	Impact data
Visual ChatGPT	Limited cultural understanding and sensitivity to symbolic nuance	75% alignment with human evaluators, compared with 90% human agreement
Flamingo	Difficulty handling ambiguity in subjective and creative tasks	Up to 15% variance in repeated evaluations
GPT-4 (Vision-enabled)	Dependence on biased input data in culturally sensitive tasks	Risk of propagating stereotypes and errors
All MLLMs	High computational cost and difficult real-time deployment	Limited scalability in constrained environments

Few-shot learning has improved visual question answering, image captioning, medical image segmentation, and cross-modal retrieval with only a few examples (Antol et al., 2015; Alayrac et al., 2022; Zhang et al., 2021a; Faghri et al., 2017). It has also been explored for subjective tasks such as art evaluation. However, artistic style and criteria vary substantially, and example selection remains difficult for rare or ambiguous cases. Existing work identifies dynamic adaptation, robust metrics, reinforcement-learning selection, and long-tail generalization as open directions (Xu et al., 2021; Li et al., 2024; Hao et al., 2019; Adiga et al., 2024; Li et al., 2023). Our focus is the multimodal, clustering-based selection of actual artwork shots.

2.3 Few-shot learning for MLLMs

Few-shot learning supplies a small number of labeled demonstrations so that an MLLM can infer the task definition and adapt to a new input. Contextual alignment and relevance-based retrieval can improve performance in specialized vision-language tasks, while diversity-focused sampling helps cover different visual concepts. In artwork evaluation, however, a demonstration can have both positive and negative effects: it may clarify the rating rubric for one image but introduce a conflicting visual pattern for another. This motivates evaluating the selection strategy itself rather than assuming that more demonstrations are always better.

3 Problem Formulation

When MLLMs rate artwork, insufficient or non-representative prompts can create a long-tail failure mode. Figure 1 illustrates the motivating observation from the supplied manuscript. With no example, the model assigns a weak score to a test image. Adding one relevant example improves the score, whereas adding another, less compatible example can reduce it. Thus, a few-shot prompt is not automatically beneficial; its effect depends on which examples are selected.

Let $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ denote the artwork identifiers. We divide the data into a shot-selection set $\mathcal{I}_S$ and a test set $\mathcal{I}_T$. Each artwork is associated with a rating vector containing nine dimensions: Realism (RE), Deformation (DE), Imagination (IM), Color Richness (CR), Color Contrast (CC), Line Combination (LC), Line Texture (LT), Picture Organization (PO), and Transformation (TR) (Biswas, 2021; Sfarra et al., 2014; Searle & Shulha, 2016; Lu et al., 2015; Pylypchuk et al., 2021; Zhang et al., 2021b; Locher et al., 1999; Ding et al., 2020; Du, 2020).

We distinguish two rating sets. S_E contains expert ratings for the test artworks, and S_M contains the ratings generated by the MLLM after the selected shots have been added to the prompt. Let W be a shot-selection function. Given $\mathcal{I}_S$ and a target number of shots k, it returns

$$X = W(\mathcal{I}_S, k), \tag{1}$$

where X is the selected set of few-shot examples. We instantiate W with random selection, K-Means selection, and K-Medoids selection.

The goal is to choose X such that the MLLM ratings agree with expert ratings on $\mathcal{I}_T$. For each rating dimension d, we compute the Pearson correlation between the corresponding components of

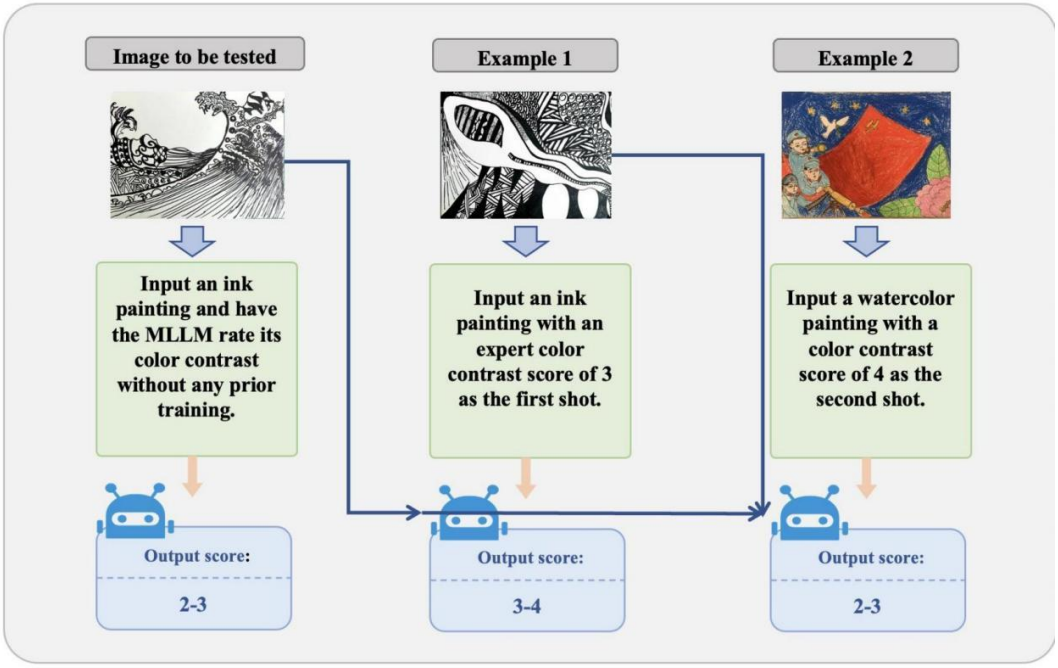


Figure 1: Example of how few-shot prompts can improve or harm an MLLM rating.

S_M and S_E:

$$r_d = \frac{\sum_{n=1}^{N}(s_{M,n,d} - \bar{s}_{M,d})(s_{E,n,d} - \bar{s}_{E,d})}{\sqrt{\sum_{n=1}^{N}(s_{M,n,d} - \bar{s}_{M,d})^2}\sqrt{\sum_{n=1}^{N}(s_{E,n,d} - \bar{s}_{E,d})^2}}. \tag{2}$$

The coefficient ranges from -1 to 1 and reflects the consistency of the MLLM's scoring trend with the expert trend. We evaluate r_d for zero-shot, one-shot, three-shot, and five-shot prompts.

4 Methodology

4.1 Clustering-based sample selection

The selection problem is to find representative artwork examples from a larger collection. Random selection is simple and independent of the empirical distribution, but it may miss important visual patterns. K-Means and K-Medoids instead use image features to partition the selection set into k clusters and choose one representative per cluster. The selected examples become the training context for the MLLM.

K-Means selection. Let $C = \{C_1, \ldots, C_k\}$ be the clusters and c_i the centroid of cluster i. K-Means minimizes the sum of squared Euclidean distances:

$$J(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, c_i)^2. \tag{3}$$

The centroid and distance are

$$c_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \qquad d(x, c_i) = \sqrt{\sum_{\ell=1}^{n} (x_\ell - c_{i\ell})^2}. \tag{4}$$

The procedure initializes k centers, alternates nearest-center assignment and centroid updates, and stops when $\|C^{(t+1)} - C^{(t)}\| < \epsilon$ or the iteration limit is reached. The image nearest to each

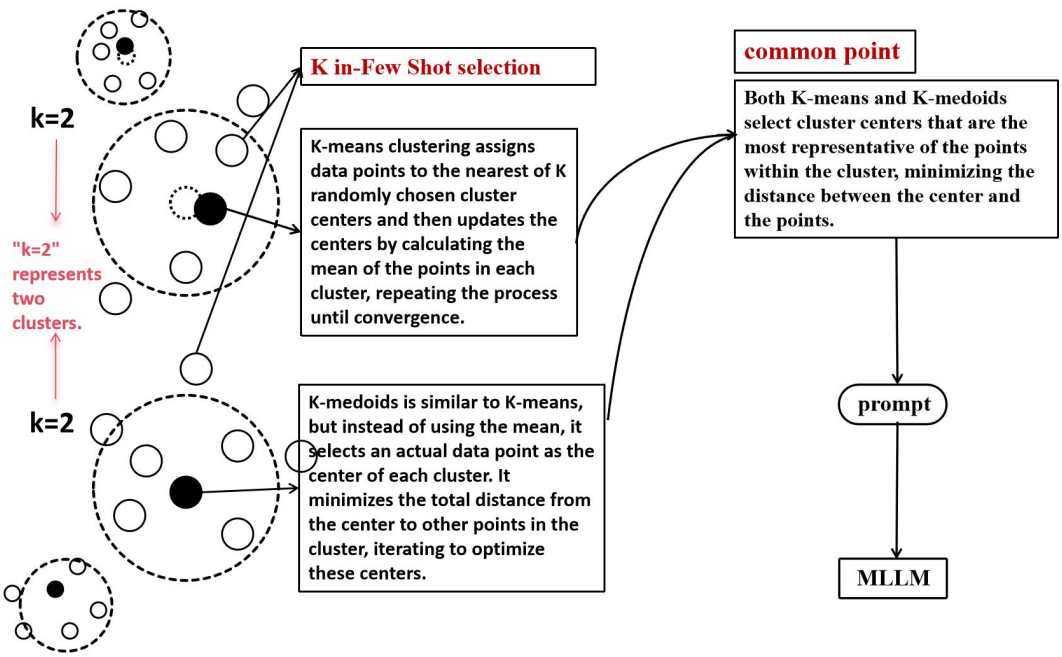


Figure 2: Comparison of K-Means and K-Medoids clustering for selecting few-shot examples.

final centroid is returned as a shot. The supplied implementation additionally applies PCA for visualization and exports assignments and centers.

K-Medoids selection. K-Medoids has the same assignment structure but chooses each center from the observed data points. Given medoids $M = \{m_1, \ldots, m_k\}$, it minimizes

$$J(M) = \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, m_i), \tag{5}$$

where the new medoid is

$$m_i = \arg\min_{x_j \in C_i} \sum_{x_t \in C_i} d(x_j, x_t), \qquad d(x, y) = \sqrt{\sum_{\ell=1}^{n} (x_\ell - y_\ell)^2}. \tag{6}$$

The algorithm alternates assignment and medoid updates until convergence, and returns the actual artwork at each medoid as a shot. Because a medoid cannot be moved to an implausible mean outside the data, this strategy is expected to be less sensitive to noise and outliers than K-Means.

Figure 2 gives the visual comparison used in the original study. Both clustering methods choose representative centers, but K-Means uses a virtual mean while K-Medoids uses an observed point. The final prompt is formed from the selected shots and passed to the MLLM for rating.

4.2 Evaluation metrics

For each selection algorithm and each value of k, we create one-shot, three-shot, and five-shot groups. The MLLM receives the selected artwork-rating pairs and then evaluates the test image set. We compare the resulting S_M with S_E using Equation 2. A high positive coefficient indicates that the MLLM follows the experts' relative scoring trend for a dimension; a coefficient near zero indicates little linear agreement, and a negative value indicates an inverse trend. Comparing coefficients across shot counts reveals whether the model benefits from additional examples and whether a selection method is sensitive to a particular dimension.

Table 3: Experimental results of random shot selection.

Shots	RE	DE	IM	CR	CC	LC	LT	PO	TR
Zero-shot	0.2964	0.0437	0.5255	0.1524	0.4140	0.2606	0.0563	-0.0112	-0.0053
One-shot	0.3192	0.0740	0.5466	0.0062	0.4081	0.3301	0.1853	-0.0112	-0.0738
Three-shot	0.2136	0.1022	0.6254	0.2353	0.4681	0.1459	-0.0875	0.0112	-0.0283
Five-shot	0.1309	0.0404	0.4170	0.2489	0.5128	0.1656	0.1303	-0.1518	0.0101

Table 4: Experimental results of K-Means-based shot selection.

Shots	RE	DE	IM	CR	CC	LC	LT	PO	TR
Zero-shot	0.2636	0.0788	0.5008	0.1734	0.4009	0.2310	0.0867	-0.0112	-0.0677
One-shot	0.1900	-0.0783	0.4097	0.2586	0.3389	0.0952	0.1408	0.2230	-0.0124
Three-shot	0.1913	0.0763	0.6637	0.1267	0.3842	0.2310	0.1190	0.3115	0.0282
Five-shot	0.2429	0.1686	0.5190	0.3141	0.5128	0.1656	0.1727	0.2668	-0.0956

5 EXPERIMENTS

The dataset consists of artwork submissions from students at several elementary schools. Experts rated each artwork on the nine dimensions defined in Section 3. We removed samples with missing, incomplete, or obviously erroneous ratings through multiple data-cleaning rounds. The cleaned data were organized into artwork-rating pairs for shot selection and testing. In total, ratings from several experts were available for 337 artworks.

We compare random, K-Means, and K-Medoids selection under zero-shot, one-shot, three-shot, and five-shot settings. The selected examples are supplied to ChatGPT-4 programmatically, after which the model rates the test artworks. Finally, we compare ChatGPT-4 ratings with the expert ratings using dimension-wise Pearson correlations. The purpose of this experiment is to isolate the effect of the shot-selection strategy while keeping the evaluated MLLM and rating dimensions fixed.

6 RESULTS

6.1 RANDOM SELECTION

Random selection improves some dimensions but is unstable across shot counts. For Imagination, the coefficient increases from 0.5255 in the zero-shot setting to 0.6254 with three shots. Color Richness increases from 0.1524 to 0.2489 between zero and five shots. In contrast, Realism increases slightly from 0.2964 to 0.3192 with one shot, then decreases to 0.2136 and 0.1309 with three and five shots. These changes indicate that random examples can help when they happen to be compatible with a dimension, but they do not reliably provide representative coverage.

6.2 K-MEANS SELECTION

K-Means performs well on Color Contrast, where the coefficient increases from 0.4140 with no shots to 0.5128 with five shots. However, Line Combination decreases from 0.2606 to 0.1656 over the same range. The method can therefore capture useful global structure when clusters are well behaved, but the mean-based center can be influenced by outliers or by broad clusters containing multiple visual styles.

6.3 K-MEDOIDS SELECTION

K-Medoids produces its largest improvement on Color Richness, increasing from 0.1459 in the zero-shot condition to 0.3492 with five shots. Picture Organization also improves from -0.0112 to 0.0918. The Realism coefficient is highest with one carefully selected shot (0.2803), whereas adding more shots lowers it to 0.1550 and 0.1952 for three and five shots, respectively. This result reinforces that shot quality can matter more than shot count for a particular rating dimension.

Table 5: Experimental results of K-Medoids-based shot selection.

Shots	RE	DE	IM	CR	CC	LC	LT	PO	TR
Zero-shot	0.2507	0.1208	0.4962	0.1459	0.4497	0.2310	0.1190	-0.0112	0.0476
One-shot	0.2803	-0.0139	0.2962	0.2137	0.3051	0.1966	0.1408	0.0112	-0.1518
Three-shot	0.1550	0.1425	0.3197	0.2014	0.4598	0.3808	0.0563	0.0161	0.0605
Five-shot	0.1952	-0.0408	0.2959	0.3492	0.5162	0.1351	-0.1488	0.0918	0.0810

6.4 Comparison

Across the three strategies, K-Medoids provides the most stable coefficients across dimensions and is particularly competitive on Color Richness and Picture Organization. This behavior is consistent with the use of observed data points as centers: the selected shots remain plausible artworks even when the feature space contains noise or outliers. Random selection exhibits larger fluctuations, while K-Means can be effective in clean clusters but is more sensitive to atypical points.

Line Texture is difficult for all three strategies. Its coefficients are generally low, and the K-Means and K-Medoids values decrease as the number of shots increases. This suggests that Line Texture may be less responsive to the current few-shot prompt or that the expert ratings for this dimension require visual and stylistic cues not captured by the selected features. Future experiments should examine richer image representations, dimension-specific retrieval, and repeated runs with confidence intervals rather than relying on a single coefficient per condition.

7 Conclusion

We reformulated automated artwork evaluation as a shot-selection problem for MLLMs and compared random, K-Means, and K-Medoids strategies. On a dataset of 337 student artworks, K-Medoids generally yields more stable alignment with expert ratings, with notable gains for Color Richness and Picture Organization. The results also show that adding demonstrations is not uniformly helpful: the optimal shot count depends on the rating dimension, and a single representative example can outperform larger prompts for some criteria.

These findings support representative and outlier-resistant retrieval as a practical component of M-LLM artwork evaluation. The approach could help provide scalable feedback in settings with limited art-education resources, while expert oversight remains important for culturally and contextually nuanced judgments. Future work should combine complementary selection strategies, evaluate more robust visual embeddings, report uncertainty across repeated selections, and investigate the low sensitivity of Line Texture.

AI use statement

In preparing this formatting revision, generative AI tools were used to convert the supplied PDF into the ICLR 2027 LaTeX layout, repair PDF-induced line-break artifacts, and format the figures, tables, equations, and references. They were not used in this revision to generate new data, formulate scientific claims, design experiments, implement the evaluated method, or interpret the reported results. The authors reviewed the resulting manuscript and remain responsible for the final content. This statement covers the formatting revision; the authors should amend it to reflect any AI use during the underlying research or original manuscript preparation and should ensure that the same disclosure is entered in the OpenReview form.

Ethics statement

The study uses artwork submitted by elementary-school students. Before submission, the authors should document data provenance, consent or applicable exemption, de-identification, and safeguards for any personally identifiable information in the artwork or metadata. The supplied manuscript does not provide those details.

Reproducibility statement

The manuscript specifies the dataset size, rating dimensions, three shot-selection strategies, clustering objectives, and Pearson-correlation evaluation. The supplied manuscript does not include the source data, exact prompts, image features, random seeds, implementation, or selected-shot files. These materials should be described or released anonymously as supplementary material so that the reported comparisons can be reproduced.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Aleman, Diogo Almeida, Janko Altenschmidt, et al. Gpt-4 technical report. Technical report, OpenAI, 2023. arXiv preprint arXiv:2303.08774.

R. Adiga, L. Subramanian, and V. Chandrasekaran. Designing informative metrics for few-shot example selection, 2024. Manuscript.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Saeid Samangooei, Marlow Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Rodrigo Barreira, Oriol Vinyals, Andrew Zisserman, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

M. Biswas. Realism. *BioScope: South Asian Screen Studies*, 12(1-2):158–161, 2021.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. Technical report, Stanford University, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ngoc Minh Bui and Jessie S. Barrot. Chatgpt as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, pp. 1–18, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

Y. Du. Research on the transformation and innovation of visual art design form based on digital fusion technology. *Applied Mathematics and Nonlinear Sciences*, 2020.

Fartash Faghri, David J. Fleet, Jamie R. Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2017. arXiv preprint arXiv:1707.05612.

Fusheng Hao, Fumin He, Jiajun Cheng, et al. Collect and select: Semantic alignment metric learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8460–8469, 2019.

J. Kotia, A. Kotwal, and R. Bharti. Few-shot learning for medical imaging. In *Machine Learning Algorithms for Industrial Applications*, pp. 107–132, 2021.

X. Li, X. Yang, and Z. Ma. Deep metric learning for few-shot image classification: A review of recent developments. *Pattern Recognition*, 138:109381, 2023.

X. Li, Y. Yang, J. Wu, et al. Mamba-fscil: Dynamic adaptation with selective state space model for few-shot class-incremental learning, 2024. Manuscript.

Paul J. Locher, Pieter J. Stappers, and Kees Overbeeke. An empirical evaluation of the visual rightness theory of pictorial composition. *Acta Psychologica*, 103:261–280, 1999.

Peng Lu, Zhaozheng Kuang, Xiaoyong Peng, et al. Discovering harmony: A hierarchical colour harmony model for aesthetics assessment. In *Computer Vision - ACCV 2014*, pp. 452–467, 2015.

O. Pylypchuk, A. Polubok, O. Krivenko, et al. Developing an approach to colour assessment of works of art on aim to creating a comfortable and harmonious interior. In *2021 International Conference on Social Sciences and Big Data Application*, pp. 181–187, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831, 2021.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, et al. Meta-learning for semi-supervised few-shot classification, 2018. Manuscript.

Michael J. Searle and Lyn M. Shulha. Capturing the imagination: Arts-informed inquiry as a method in program evaluation. *Canadian Journal of Program Evaluation*, 31(1-2):34–60, 2016.

Stefano Sfarra, Clemente Ibarra-Castanedo, Dario Ambrosini, et al. Discovering the defects in paintings using non-destructive testing techniques and passing through measurements of deformation. *Journal of Nondestructive Evaluation*, 33:358–383, 2014.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chenfei Wu, Sheng Yin, Weikang Qi, et al. Visual chatgpt: Talking, drawing and editing with visual foundation models, 2023. arXiv preprint arXiv:2303.04671.

Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*, pp. 2397–2406, 2016.

Chen Xu, Yanwei Fu, Chen Liu, et al. Learning dynamic alignment via meta-filter for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5182–5191, 2021.

Shuyang Yu, Rui Bao, Parminder Bhatia, et al. Dynamic uncertainty ranking: Enhancing in-context learning for long-tail knowledge in llms, 2024. Manuscript.

Guowei Zhang, Guang Kang, Yi Yang, et al. Few-shot segmentation via cycle-consistent transformer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21984–21996, 2021a.

J. Zhang, Y. Miao, and J. Yu. A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges. *IEEE Access*, 9:77164–77187, 2021b.

C. Zheng, Z. Yu, Y. Jiang, et al. Artmentor: Ai-assisted evaluation of artworks to explore multimodal large language models capabilities. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2025.

L. Zhu, T. Chen, D. Ji, et al. Llafs: When large language models meet few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3065–3075, 2024.