Can Large Language Models Understand Visual Aesthetics for Intelligent Image Editing

Anshu Raj
Drawify
Nolensville, USA

***Abstract*— Recent progress in multimodal AI has significantly expanded the capabilities of automated image editing systems. Contemporary models can perform a variety of complex editing operations, including object manipulation, style transformation, and lighting adjustment. However, despite these advancements, many existing systems still struggle to capture subjective aspects of visual aesthetics, such as compositional quality, visual balance, and stylistic harmony. These elements play a crucial role in determining how viewers perceive and evaluate visual content, yet they remain difficult to model computationally. In this work, we investigate whether large language models can reason about visual aesthetics and provide meaningful guidance for intelligent image editing. We propose a framework in which a language model analyzes an input image and generates structured aesthetic critiques together with actionable editing recommendations. These recommendations are then used to guide an image editing module that performs targeted modifications, including lighting enhancement, object repositioning, background simplification, and color adjustment.**

***Index Terms*— Image Editing, Large Language Models, Multimodal AI, Visual.**

I. Introduction

Recent developments in AI have substantially advanced the capabilities of automated image editing technologies [1]. Modern computer vision systems can now perform a wide variety of editing operations, including object removal [2], background replacement [3], style transformation [4], and lighting adjustment [5]. Many of these capabilities are supported by diffusion-based generative models and multimodal architectures that enable sophisticated manipulation of visual content while maintaining a high degree of visual realism. Tools powered by models such as Stable Diffusion [6] and DALL·E [7] allow users to modify images using natural language instructions, making advanced image editing more accessible to individuals without specialized technical expertise. Despite these technological advances, most automated editing systems remain primarily focused on preserving realism and semantic correctness rather than improving the overall aesthetic quality of images.

Visual aesthetics plays a central role in how images are perceived and interpreted. The perceived quality of an image is often influenced by multiple factors, including composition, lighting, color harmony, visual balance, and stylistic consistency. Professional photographers, designers, and visual artists routinely evaluate images using these principles when deciding how to refine elements such as exposure, contrast, framing, or object placement. However, translating such aesthetic judgments into computational systems remains challenging. Aesthetic perception is inherently subjective and may depend on contextual, cultural, and stylistic considerations. However, many automated image editing tools apply generalized enhancement techniques without determining whether these adjustments genuinely improve the visual appeal or communicative clarity of the image.

Recent progress in Large Language Models (LLMs) has opened new possibilities for integrating reasoning capabilities into visual processing tasks. Advanced models such as GPT-4 [8] have demonstrated strong performance in interpreting complex instructions, generating structured explanations, and reasoning across multimodal inputs. These capabilities suggest that language models may be capable of analyzing images through descriptive reasoning and identifying aesthetic properties that influence visual perception. Instead of relying solely on traditional image processing algorithms, a language-based system could evaluate an image, detect potential aesthetic limitations, and propose editing recommendations similar to those suggested by human designers—for example, improving lighting balance, reducing visual clutter, or adjusting compositional framing. Integrating language-based reasoning with automated image editing systems introduces a new paradigm for intelligent visual enhancement. Within this framework, a language model serves as an aesthetic reasoning module that generates critiques and editing recommendations informed by established visual design principles. These recommendations can then guide downstream editing modules responsible for applying specific transformations to the image. Such an approach offers several potential benefits, including increased transparency in editing decisions, closer alignment with human aesthetic preferences, and the ability to incorporate high-level design knowledge into automated editing workflows.

This study investigates the potential role of LLMs in supporting intelligent image editing through aesthetic reasoning. Specifically, we examine how language models can analyze visual content, produce structured aesthetic critiques, and generate editing instructions that guide automated image enhancement processes. In addition, we review related research in computational aesthetics, multimodal reasoning, and AI-driven image editing to better understand the current state of the field.

The study is organized as follows: Section II presents Background (Foundations of Visual Aesthetics, Computational Aesthetics, and Editing Advances); Section III discusses Large Language Models for Multimodal Aesthetic Reasoning; Section IV introduces the Language-Guided Intelligent Image Editing Framework; Section V covers Datasets for Aesthetic Evaluation and Image Editing; Section VI explains Evaluation Metrics; Section VII discusses Open Challenges; Section VIII presents Future Research Directions; and Section IX concludes the study.

II. Background

A. Foundations of Visual Aesthetics

The concept of visual aesthetics is grounded in well-established principles from art theory and design practice [9]. Designers and photographers commonly rely on compositional guidelines to organize visual elements within an image in a structured and meaningful way. Principles such as balance, contrast, emphasis, and spatial harmony are widely used to guide viewer attention and enhance visual communication. For example, balanced compositions distribute visual weight evenly across an image so that no single region disproportionately dominates the frame. Contrast is often employed to highlight important subjects by distinguishing them from surrounding elements, while harmonious color combinations help maintain visual coherence and prevent distracting visual conflicts. In professional design environments, these principles are rarely applied in a purely mechanical way. Instead, designers combine formal guidelines with creative judgment and contextual awareness. Factors such as narrative intention, cultural interpretation, and stylistic preferences often influence how aesthetic decisions are made. Because aesthetic perception is shaped by both objective visual properties and subjective interpretation, modeling these principles within computational systems remains a complex and challenging task.

B. Computational Aesthetics

The field of computational aesthetics seeks to model and predict aesthetic perception using algorithmic and data-driven approaches [10]. Early research in this area focused on handcrafted visual features designed to approximate established photographic and artistic principles. Researchers developed algorithms capable of analyzing image attributes such as color distributions, texture patterns, brightness variations, and compositional arrangements. These features were then used to train machine learning models that could estimate aesthetic quality or rank images according to perceived visual appeal. Progress in this field accelerated with the introduction of large-scale annotated datasets that contain aesthetic ratings provided by human observers. One of the most widely used datasets is the AVA Dataset [11], which includes hundreds of thousands of photographs accompanied by aesthetic ratings collected from online photography communities. Training deep learning models on such datasets has enabled researchers to learn complex visual representations that correlate more closely with human aesthetic preferences. In particular, convolutional neural networks and other deep learning architectures have demonstrated substantial improvements in predicting aesthetic quality when compared with earlier feature-based approaches.

C. Advances in Automated Image Editing

In parallel with research on aesthetic evaluation, automated image editing technologies have also advanced considerably [12]. Traditional image editing tools primarily focused on low-level enhancement operations [13] such as brightness adjustment, color correction, sharpening, and noise reduction. While these techniques improved image clarity, they typically required manual user intervention or relied on predefined enhancement rules. Recent developments in deep learning have significantly expanded the capabilities of automated image editing systems [14]. Modern generative models can perform complex tasks such as object removal, background modification, and artistic style transfer while preserving visual realism. Diffusion-based generative models, including Stable Diffusion [15] and DALL·E, have further broadened the possibilities for image editing by enabling users to manipulate visual content through natural language instructions. These systems can generate or modify images while maintaining semantic consistency with textual descriptions, making advanced editing capabilities more accessible to a wider range of users.

III. Large Language Models for Multimodal Aesthetic Reasoning

A. Large Language Models and Reasoning Capabilities

LLMs are generally built on transformer-based neural architectures [16]. The transformer architecture enables models to process long sequences of information and capture relationships between tokens through self-attention mechanisms. This design allows models to learn contextual dependencies across large bodies of text and generate coherent responses based on complex patterns of language use. Building upon this architectural foundation, modern language models are trained on vast and diverse datasets that include books, web pages, and other textual resources. Through this large-scale training process, they acquire knowledge of linguistic structures, factual information, and reasoning patterns. As a result, contemporary models such as GPT-4 and PaLM [17] demonstrate strong performance across a variety of tasks, including summarization, question answering, code generation, and structured reasoning. These models are capable of interpreting detailed instructions and producing step-by-step explanations that resemble human analytical reasoning. Such capabilities are particularly relevant for tasks that involve interpreting abstract concepts or generating structured plans, suggesting that language models may also be able to reason about higher-level visual attributes such as aesthetic quality and design principles.

B. Vision–Language Models and Multimodal Understanding

In addition to text-based processing, recent research has expanded language models to operate within multimodal environments that incorporate visual inputs [18]. Vision–language models [19] integrate textual and visual representations, enabling systems to analyze images and generate language-based descriptions or reasoning outputs.

These models typically learn joint representation spaces that align textual information with visual features extracted from images. A notable example is CLIP [20], which learns shared representations between images and text using contrastive learning techniques. CLIP has shown strong ability in associating textual descriptions with visual concepts and has become a foundational component in many modern image generation and editing systems.

C. Language-Based Aesthetic Reasoning

Aesthetic evaluation frequently involves describing visual qualities using natural language [21]. Photographers, designers, and art critics often express aesthetic judgments through descriptive observations such as noting that a composition lacks balance, that lighting conditions appear overly harsh, or that background elements distract from the primary subject. These forms of critique naturally combine visual interpretation with linguistic reasoning, making them well suited to language-based modeling approaches. LLMs can therefore function as effective aesthetic reasoning agents capable of analyzing visual content and generating structured critiques of image quality. When integrated with visual perception modules, these models can interpret visual features and produce explanations regarding compositional weaknesses, lighting issues, or stylistic inconsistencies. Such explanations can then be converted into actionable editing instructions, including adjustments to brightness, background simplification, object repositioning, or modifications to color schemes to enhance visual harmony. This language-guided reasoning approach offers a significant advantage over purely numerical image processing methods. Instead of applying generic enhancement operations, systems can generate context-aware editing recommendations that reflect human-like aesthetic reasoning.

IV. Language-Guided Intelligent Image Editing Framework

A. Overview of the Framework

Conventional image editing systems typically rely on predefined enhancement operations or manual adjustments specified by users. These systems generally modify visual attributes such as brightness, contrast, or color balance without explicitly determining whether such modifications actually improve the aesthetic quality of the image. The proposed framework addresses this limitation by introducing an intermediate reasoning stage that evaluates the visual characteristics of an image before any editing operations are performed (see Fig. 1). This reasoning stage allows the system to identify aesthetic shortcomings and generate targeted recommendations for improvement. The framework is composed of three main components that operate sequentially within the editing pipeline. First, a visual analysis module extracts semantic and compositional features from the input image. Next, a LLM interprets these visual representations and generates aesthetic critiques together with structured editing recommendations. Finally, an image editing module applies modifications to the image based on the guidance produced by the language model.

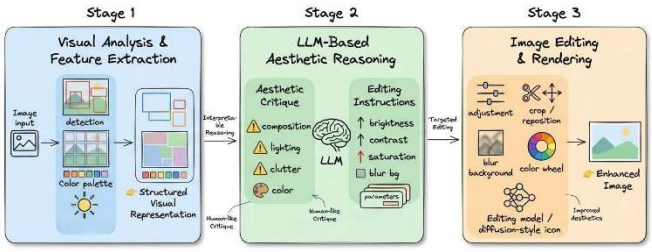

Fig.1 Language-Guided Image Editing Pipeline

B. Visual Analysis and Feature Extraction

The initial stage of the framework focuses on extracting meaningful visual information from the input image. Computer vision models analyze the image to detect objects, understand spatial relationships, evaluate lighting conditions, and identify color distributions. These extracted visual attributes provide essential contextual information required for aesthetic evaluation. Vision–language models, such as CLIP, are particularly valuable during this stage because they enable visual information to be represented in a format that can be interpreted by language-based reasoning systems. Through this process, the system generates a structured representation of the image that may include details about object placement, compositional layout, brightness patterns, and color relationships. These representations form the basis for the higher-level aesthetic reasoning performed by the language model in subsequent stages.

C. Aesthetic Critique Generation with Language Models

During the second stage of the framework, a LLM examines the visual representation of the image and generates an aesthetic critique. The model interprets visual features in the context of established design principles and produces descriptive feedback regarding the strengths and weaknesses of the image. Advanced language models are capable of generating critiques that resemble human evaluations, identifying issues such as uneven lighting, distracting background elements, or imbalanced composition. The critique produced by the language model typically includes both qualitative analysis and practical recommendations for improvement. For example, the model may recommend increasing brightness in darker regions, adjusting color saturation to enhance visual harmony, or repositioning elements within the frame to improve compositional balance. Because these critiques are expressed in natural language, they provide clear and interpretable explanations that allow users to understand the reasoning behind each suggested modification.

D. Editing Instruction Generation

After generating an aesthetic critique, the language model converts its observations into structured editing instructions that can be executed by the image editing module. This stage transforms qualitative design feedback into concrete editing actions. For instance, a recommendation to improve lighting balance may translate into specific adjustments to exposure or contrast, while a suggestion to reduce visual clutter may involve background blurring or object removal. Structured editing instructions play a crucial role in bridging the gap between high-level aesthetic reasoning and the technical operations required for image modification. Instead of

applying generalized enhancement techniques, the editing module performs targeted adjustments designed to address the particular aesthetic issues identified during the critique stage.

E. Image Editing and Rendering

In the final stage of the framework, the editing module applies the recommended modifications to the image based on the structured instructions generated by the language model. Modern image editing systems are capable of performing a wide range of transformations, including color corrections, lighting adjustments, object repositioning, and background editing. Additionally, generative models such as Stable Diffusion can be integrated into the pipeline to support more advanced modifications while preserving visual realism. Because the editing process is guided by structured instructions derived from aesthetic reasoning, the resulting images are expected to demonstrate improved visual quality compared with those produced by conventional automated enhancement techniques.

V. Datasets for Aesthetic Evaluation and Image Editing

The development of intelligent image editing systems that incorporate aesthetic reasoning relies heavily on the availability of datasets that reflect both visual content and human perceptions of aesthetic quality (see Fig. 2). Unlike conventional computer vision datasets, which primarily emphasize tasks such as object recognition or scene classification, datasets used in computational aesthetics and image editing research often include annotations related to visual appeal, compositional balance, and photographic style. One of the most widely used resources for aesthetic evaluation research is the AVA Dataset [11]. This dataset contains more than 250,000 images collected from an online photography community, where each image is associated with aesthetic ratings assigned by photography enthusiasts. These ratings reflect collective judgments regarding the visual appeal of photographs and are commonly used to train models that predict aesthetic scores or classify images based on their perceived quality. Another important dataset in this area is the Aesthetic and Attributes Database (AADB) [22], which was specifically created to support research on aesthetic prediction and attribute analysis. In addition to overall aesthetic ratings, AADB provides annotations for several aesthetic attributes, including color harmony, depth of field, and lighting conditions. These attribute-level annotations offer richer insights into the factors that influence aesthetic perception, enabling models to learn more detailed relationships between visual features and human aesthetic judgments.

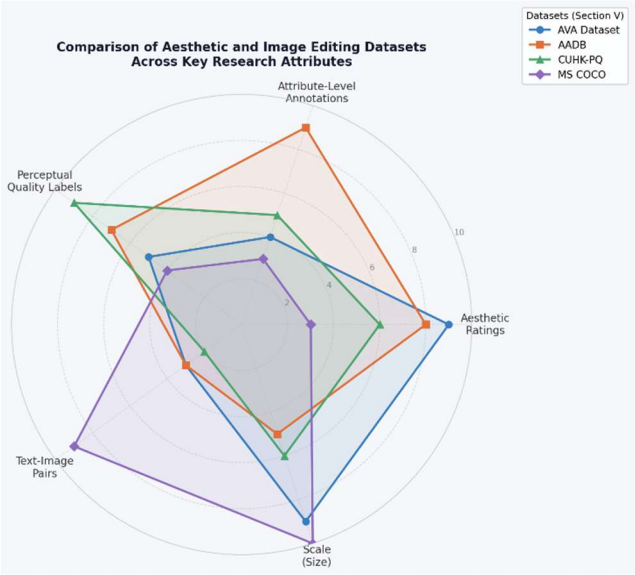

Fig.2 Comparison of aesthetic datasets across key research attributes.

Datasets that capture perceptual image quality are also relevant for studies of aesthetic analysis and editing. For instance, the CUHK-PQ Dataset [23] contains images annotated with perceptual quality ratings that represent human evaluations of visual clarity and appeal. Although the dataset was initially designed for image quality assessment tasks, it provides useful information regarding how viewers perceive factors such as contrast, structural integrity, and overall visual clarity, which are closely connected to aesthetic evaluation. Beyond datasets designed specifically for aesthetic assessment, several large-scale image datasets play an important role in research related to image generation and editing. A notable example is the MS COCO [24], which consists of images paired with descriptive textual captions. While the dataset was originally introduced for tasks such as object detection and image captioning, it has become an essential resource for multimodal learning and text-guided image editing systems. The dataset contains complex visual scenes with multiple interacting objects, making it particularly valuable for training models that interpret visual information in relation to natural language descriptions (see Table I).

TABLE I
SUMMARY OF DATASETS USED IN AESTHETIC IMAGE EDITING

Dataset	Size	Aesthetic Ratings	Attribute Annotations	Text-Image Pairs	Primary Application
AVA Dataset	255,000 images	Yes	No	No	Aesthetic score prediction
AADB	10,000 images	Yes	Yes	No	Attribute-level aesthetic analysis
CUHK-PQ	17,690 images	Partial	No	No	Perceptual image quality assessment
MS COCO	330,000 images	No	Partial	Yes	Multimodal and text-guided editing

VI. Evaluation Metrics for Aesthetic Image Quality

Assessing the performance of intelligent image editing systems requires evaluation methodologies capable of measuring both the technical quality of edited images and their perceived aesthetic improvement (see Fig. 3). Unlike traditional computer vision tasks that rely on objective accuracy metrics, aesthetic evaluation inherently involves subjective judgments influenced by human perception. For this reason, researchers often employ a combination of automated evaluation metrics and human-centered assessment methods to determine whether editing approaches successfully enhance visual appeal and compositional quality. One category of evaluation metrics focuses on image realism and distributional similarity. In generative modeling research, the Fréchet Inception Distance (FID) is widely used to measure the similarity between distributions of generated images and real images. This metric evaluates statistical differences between feature representations extracted from deep neural networks. Lower FID scores generally indicate that generated or edited images more closely resemble real images in terms of visual characteristics. Although this metric is effective for evaluating visual realism, it primarily measures distributional similarity rather than aesthetic quality. Another commonly used metric in generative image evaluation is the Inception Score (IS). This metric assesses both the diversity and recognizability of generated images based on predictions produced by an image classification model. Higher IS typically indicate that generated images contain distinct and recognizable visual structures while maintaining diversity across samples. Similar to FID, however, this metric is more closely related to image generation quality than to aesthetic composition. To evaluate structural consistency between images, researchers frequently employ the Structural Similarity Index (SSIM). SSIM measures similarity between two images by comparing luminance, contrast, and structural features. In image editing applications, this metric can be used to determine whether an edited image preserves important structural details from the original image while incorporating aesthetic improvements (see Table II).

TABLE II
PERFORMANCE COMPARISON OF EVALUATION METRICS IN AESTHETIC IMAGE EDITING SYSTEMS

Metric	**Score Range**	**Conventional System**	**LLM-Guided System**	**Aesthetic Coverage**
Frechet Inception Distance	Lower is better	64.30	38.70	Low
Inception Score	Higher is better	5.60	7.80	Low
Structural Similarity Index	0 to 1	0.63	0.84	Moderate
Visual Balance Score	0 to 5	2.90	4.30	High
Color Harmony Rating	0 to 5	3.10	4.40	High
Human Preference Score	0 to 100	54.20	83.60	Very High

Because aesthetic perception is inherently subjective, human evaluation remains one of the most reliable approaches for assessing aesthetic improvements. In many studies, participants are asked to compare original images with edited versions and indicate which version they find more visually appealing. Other evaluation methods involve rating images according to attributes such as compositional balance, color harmony, clarity of subject emphasis, and overall aesthetic quality. Human preference studies are particularly valuable when evaluating language-guided editing systems, as these systems aim to produce results that align closely with human aesthetic judgments.

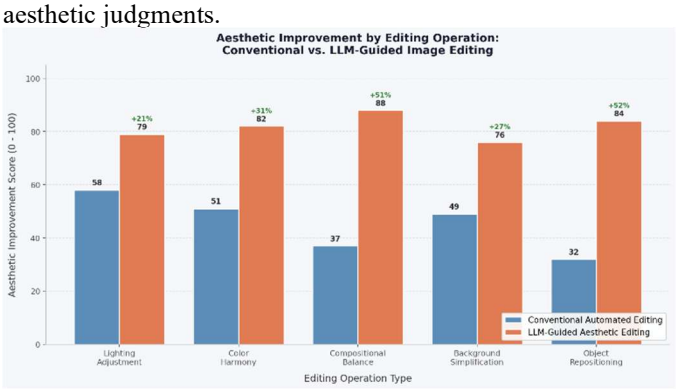

Fig. 3. Aesthetic improvement scores across editing operations: conventional vs. LLM-guided.

VII. Open Challenges in Language-Guided Aesthetic Image Editing

Despite significant advancements in multimodal AI and generative image editing technologies, a number of challenges remain in developing systems capable of understanding and enhancing visual aesthetics through language-based reasoning. Although LLMs demonstrate strong capabilities in interpreting instructions and generating structured explanations, translating these abilities into reliable aesthetic reasoning for image editing introduces several conceptual and technical complexities. Addressing these issues is essential for building automated systems that can perform image editing tasks in ways that align with human aesthetic preferences.

One of the most fundamental challenges arises from the inherently subjective nature of aesthetic perception. Unlike tasks such as object recognition or scene classification, aesthetic judgment varies widely across individuals, cultural contexts, and artistic traditions. An image that appears visually pleasing to one observer may not evoke the same response from another. This variability makes it difficult to define universal standards of aesthetic quality that can be consistently applied across diverse images and use cases. Although language models trained on large-scale data may capture general patterns associated with aesthetic evaluation, ensuring that their reasoning reflects diverse perspectives and cultural interpretations remains a significant research challenge. Another important difficulty involves representing aesthetic knowledge within computational systems. Many design principles—such as balance, harmony, contrast, and compositional flow—are typically expressed through qualitative descriptions rather than precise mathematical formulations. Human designers often rely on experience,

intuition, and contextual interpretation when applying these principles.

Integrating language-based reasoning with visual editing systems also presents a major technical challenge. While language models are capable of generating detailed critiques and editing suggestions, these recommendations must ultimately be translated into concrete image manipulation operations. Mapping natural language descriptions of aesthetic improvements to specific editing actions—such as adjusting lighting, modifying color balance, repositioning visual elements, or simplifying backgrounds—requires careful coordination between language reasoning modules and visual editing algorithms. Ensuring that the final editing operations accurately reflect the intent of the language-generated critique is an important challenge for system design. The availability of suitable training data also represents a significant limitation for research in language-guided aesthetic editing. Many existing datasets focus primarily on aesthetic scoring or object-level annotations related to visual recognition tasks. Only a limited number of datasets include examples that directly link images with aesthetic critiques and corresponding editing actions. Without such data, it becomes difficult to train models that can generate meaningful editing recommendations. The development of datasets containing image–critique pairs or structured aesthetic feedback could therefore play a crucial role in advancing this area of research.

VIII. Future Research Directions

The integration of LLMs with automated image editing technologies opens several promising avenues for future research. One promising research direction involves the design of evaluation metrics that more accurately capture human perceptions of aesthetic quality. Existing evaluation metrics primarily focus on measuring image fidelity or structural similarity, but they often fail to account for visual attributes such as compositional balance, stylistic harmony, or narrative clarity. Future research may explore computational models capable of estimating these aesthetic attributes more reliably. Combining such models with human evaluation studies could lead to more comprehensive frameworks for assessing the effectiveness of aesthetic image editing systems. Finally, improvements in vision–language understanding are likely to further enhance the ability of language models to interpret visual content and reason about aesthetic properties. As multimodal models become increasingly capable of analyzing images and generating structured explanations, they may serve as powerful tools for evaluating and refining visual designs.

IX. Conclusion

This study examined the potential role of LLMs in enabling intelligent image editing through aesthetic reasoning. Owing to their ability to interpret complex instructions and generate structured explanations, language models can analyze visual content and produce critiques that resemble human aesthetic feedback. These critiques can then be transformed into actionable editing recommendations that guide image editing systems toward improvements in areas such as lighting conditions, compositional balance, color relationships, and visual clarity. Integrating language-based reasoning with visual editing modules therefore represents a promising approach for developing automated systems that align more closely with human perceptions of aesthetic quality. In addition, this study reviewed key developments across several related research areas, including computational aesthetics, multimodal vision–language modeling, and generative image editing technologies. The discussion highlighted how these domains intersect in the development of language-guided aesthetic editing systems. The study also examined commonly used datasets for aesthetic evaluation and image editing research, outlined evaluation metrics used to measure improvements in visual quality, and discussed several open challenges associated with aesthetic subjectivity, knowledge representation, and the integration of multimodal reasoning systems. These challenges underscore the inherent complexity of modeling aesthetic judgment within computational frameworks.

References

[1] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and V. Gool, "RePaint: Inpainting Using Denoising Diffusion Probabilistic Models," *Thecvf.com*, pp. 11461–11471, 2022, Accessed: Mar. 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Lugmayr_RePaint_Inpainting_Using_Denoising_Diffusion_Probabilistic_Models_CVPR_2022_paper.html

[2] L. Rout, A. Parulekar, C. Caramanis, and S. Shakkottai, "A Theoretical Justification for Image Inpainting using Denoising Diffusion Probabilistic Models," *arXiv.org*, 2023. https://arxiv.org/abs/2302.01217 (accessed Mar. 21, 2026).

[3] O. Avrahami, D. Lischinski, and O. Fried, "Blended Diffusion for Text-Driven Editing of Natural Images," *Thecvf.com*, pp. 18208–18218, 2022, Accessed: Mar. 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Avrahami_Blended_Diffusion_for_Text-Driven_Editing_of_Natural_Images_CVPR_2022_paper.html?ref=https://githubhelp.com

[4] M. Kwon, J. Jeong, and Y. Uh, "Diffusion Models already have a Semantic Latent Space," *arXiv.org*, 2022. https://arxiv.org/abs/2210.10960 (accessed Mar. 21, 2026).

[5] M. Afifi, Derpanis, Konstantinos G, B. Ommer, and M. S. Brown, "Learning Multi-Scale Photo Exposure Correction," *Thecvf.com*, pp. 9157–9167, 2021, Accessed: Mar. 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Afifi_Learning_Multi-Scale_Photo_Exposure_Correction_CVPR_2021_paper.html

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," *Thecvf.com*, pp. 10684–10695, 2022, Accessed: Mar. 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html?utm_source=rns.dwaiai.de

[7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Openai, "Hierarchical Text-Conditional Image Generation with CLIP Latents," 2022. Available: https://3dvar.com/Ramesh2022Hierarchical.pdf

[8] penAI *et al.*, "GPT-4 Technical Report," *arXiv.org*, 2023. https://arxiv.org/abs/2303.08774 (accessed Mar. 21, 2026).

[9] M. Daryanavard Chounchenani, A. Shahbahrami, R. Hassanpour, and G. Gaydadjiev, "Deep Learning Based Image Aesthetic Quality Assessment- A Review," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–36, Feb. 2025, doi: https://doi.org/10.1145/3716820.

[10] S. K. Ray *et al.*, "Do Clinical Question Answering Systems Really

Need Specialised Medical Fine Tuning?," *arXiv.org*, 2026. https://arxiv.org/abs/2601.12812 (accessed Mar. 21, 2026).

[11] V. Nieto, L. Celona, and F. Labrador, "Understanding Aesthetics with Language: A Photo Critique Dataset for Aesthetic Assessment," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34148–34161, Dec. 2022, Accessed: Mar. 21, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/dcd18e50ebca0af89187c6e35dabb584-Abstract-Datasets_and_Benchmarks.html

[12] J.-O. Kropp, C. Schiffer, K. Amunts, and T. Dickscheid, "Denoising Diffusion Probabilistic Models for Image Inpainting of Cell Distributions in The Human Brain," *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, May 2024, doi: https://doi.org/10.1109/isbi56570.2024.10635384.

[13] A. Ignatov, V. Gool, and R. Timofte, "Replacing Mobile Camera ISP With a Single Deep Learning Model," *Thecvf.com*, pp. 536–537, 2020, Accessed: Mar. 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w31/Ignatov_Replacing_Mobile_Camera_ISP_With_a_Single_Deep_Learning_Model_CVPRW_2020_paper.html

[14] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023, doi: https://doi.org/10.1109/tpami.2023.3261988.

[15] J. Zhang, Q. Huang, J. Liu, X. Guo, and D. Huang, "Diffusion-4K: Ultra-High-Resolution Image Synthesis with Latent Diffusion Models," *Thecvf.com*, pp. 23464–23473, 2025, Accessed: Mar. 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Diffusion-4K_Ultra-High-Resolution_Image_Synthesis_with_Latent_Diffusion_Models_CVPR_2025_paper.html

[16] S. Tripathi, M. T. Nafis, I. Hussain, and J. Gao, "The Confidence Paradox: Can LLM Know When It's Wrong?," *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2078–2087, 2025, doi: https://doi.org/10.18653/v1/2025.ijcnlp-long.113.

[17] R. Anil *et al.*, "PaLM 2 Technical Report," *arXiv.org*, 2023. https://arxiv.org/abs/2305.10403 (accessed Mar. 21, 2026).

[18] J.-B. Alayrac *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, Dec. 2022, Accessed: Mar. 21, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html

[19] R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models," *arXiv.org*, 2021. https://arxiv.org/abs/2108.07258 (accessed Mar. 21, 2026).

[20] V. Govindarajan, P. Patel, S. Tripathi, M. A. Hoque, and G. S. Kashyap, "MAGIC-Enhanced Keyword Prompting for Zero-Shot Audio Captioning with CLIP Models," *arXiv.org*, 2025. https://arxiv.org/abs/2509.12591 (accessed Mar. 21, 2026).

[21] H. Yang, Y. Li, X. Jin, X. Zhou, P. Shi, and Y. Liu, "Aesthetic multi-attributes network for image captioning," *Computers and Electrical Engineering*, vol. 123, p. 110103, Apr. 2025, doi: https://doi.org/10.1016/j.compeleceng.2025.110103.

[22] J. Wu and D. Li, "Digital image quality evaluation based on multi-scale aesthetic features and graph convolutional neural networks," *Scientific Reports*, vol. 15, no. 1, Dec. 2025, doi: https://doi.org/10.1038/s41598-025-29226-5.

[23] M. Zhou *et al.*, "Blind Image Quality Assessment: Exploring Content Fidelity Perceptibility via Quality Adversarial Learning," *International Journal of Computer Vision*, vol. 133, no. 6, pp. 3242–3258, Jan. 2025, doi: https://doi.org/10.1007/s11263-024-02338-7.

[24] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *Lecture Notes in Computer Science*, pp. 740–755, 2014, doi: https://doi.org/10.1007/978-3-319-10602-1_48.