When Language Meets Visual Design Can Large Language Models Improve Automated Image Composition

Anshu Raj
Drawify
Nolensville, USA

***Abstract*— Automated image generation systems have achieved remarkable progress in producing visually appealing images. However, these systems often lack an explicit understanding of fundamental visual design principles, such as balance, hierarchy, contrast, and spatial harmony. As a result, the generated compositions may appear visually plausible but frequently exhibit inconsistencies in aesthetic structure and layout organization. Addressing this limitation requires integrating higher-level reasoning about design principles into the image generation process. In this study, we explore whether large language models can support automated image composition by reasoning about visual design concepts expressed through natural language. We demonstrate that incorporating language-driven reasoning into image composition pipelines can substantially enhance both the interpretability and aesthetic quality of automatically generated designs.**

***Index Terms*— Designs, Generation Systems, Images, Large Language Models.**

I. Introduction

Recent advances in generative artificial intelligence have significantly transformed the field of automated image creation [1]. Modern generative systems, particularly those built on diffusion models and multimodal transformer architectures, are capable of producing highly realistic images from textual descriptions. These technologies have enabled a wide range of applications, including artistic content generation [2], automated illustration [3], advertising design [4], and social media graphics [5]. Models such as DALL·E [6], Stable Diffusion [7], and Midjourney [8] have demonstrated remarkable ability to synthesize visually detailed images that align closely with user-provided prompts. Despite these impressive developments, current image generation systems continue to face challenges in an essential aspect of visual communication. Furthermore, image composition refers to the deliberate arrangement of visual elements within a frame in order to achieve clarity, aesthetic balance, and effective communication of ideas [9]. However, most contemporary generative image models primarily prioritize semantic alignment between textual prompts and visual output. While this allows them to generate objects and scenes that correspond to the given description, they often lack the structured reasoning necessary to enforce layout constraints or design-oriented composition. Therefore, generated images may appear visually plausible yet still exhibit disorganized layouts, inconsistent emphasis, or limited aesthetic coherence.

In recent years, Large Language Models (LLMs) have emerged as powerful systems for structured reasoning [10], knowledge representation [11], and instruction generation [12]. Advanced language models such as GPT-4 [13] and PaLM [14] have demonstrated strong capabilities in interpreting complex instructions, synthesizing conceptual knowledge, and generating structured plans through natural language. These abilities suggest an intriguing possibility that combining language-based reasoning with visual generation mechanisms offers a promising direction for improving the controllability and interpretability of automated design systems. Rather than generating images directly from textual prompts alone, a language model can first analyze the intended design objectives—for example, emphasizing a focal object or maintaining balanced spatial organization—and convert these goals into structured composition directives. Such directives may include layout constraints, object placement guidance, hierarchical emphasis rules, and alignment specifications. These instructions can then be used to guide downstream visual generation components, including diffusion-based image generators or layout engines. Within this framework, the language model acts as an intermediary layer that connects human design knowledge with machine-driven image synthesis.

Therefore, this study investigates the potential role of LLMs in enhancing automated image composition through language-driven reasoning about visual design principles. Specifically, we examine how natural language representations of design guidelines can be interpreted and transformed into structured composition instructions that guide image generation or editing systems. The central hypothesis of this work is that incorporating language-mediated reasoning into image generation pipelines can improve both the interpretability and the aesthetic coherence of automatically generated visual designs.

The study is organized as follows: Section II presents Background; Section III covers LLM-Guided Image Composition Framework; Section IV discusses Datasets and Benchmarks for Image Composition and Design Generation; Section V explains Evaluation Metrics for Aesthetic and Layout Quality; Section VI highlights Open Challenges in

Language-Guided Image Composition; Section VII outlines Future Research Directions; and Section IX concludes the study.

II. Background

A. Fundamentals of Visual Composition

Image composition refers to the spatial organization of visual elements—such as objects, text, color regions, and graphical shapes—within a visual frame [15]. The primary goal of composition is to guide the viewer's attention toward key elements while maintaining harmony across the entire layout. Classical design theory emphasizes that an effective composition directs visual flow in a manner that feels intuitive and structured to the viewer. One commonly referenced guideline in visual composition is the rule of thirds, which divides an image into a 3 × 3 grid and suggests positioning important elements along the grid lines or at their intersections. This approach is widely used in photography and visual media to create balanced and engaging images. Another essential concept is visual balance, which refers to the distribution of visual weight across a composition so that no single region appears excessively dominant or neglected. In design contexts such as posters, advertisements, or social media graphics, composition must also consider the relationship between textual and visual elements. Designers carefully manage spacing, alignment, and grouping to ensure that textual information remains readable while integrating naturally with surrounding visual components. Achieving this balance is essential for maintaining both functional clarity and aesthetic appeal. Although such compositional principles are well understood and consistently applied by human designers, they are rarely incorporated explicitly into modern generative image models. Instead, most automated image generation systems rely on large datasets to learn statistical associations between textual prompts and visual features. As a result, generated images may contain detailed visual content but still fail to adhere to compositional structures that a human designer would naturally implement.

B. Core Principles of Visual Design

Visual design theory identifies several foundational principles that influence the effectiveness of a composition [16]. These principles are widely taught in design education and serve as practical guidelines in professional creative workflows. One fundamental principle is balance, which concerns the distribution of visual weight across a composition. Balanced designs may be symmetrical, where elements mirror each other across a central axis, or asymmetrical, where different elements contribute varying visual weights that nonetheless create a stable overall structure. Another important principle is visual hierarchy, which determines the sequence in which viewers perceive elements within a design. Designers typically establish hierarchy by manipulating size, color, contrast, or position so that the most important elements attract attention first. This hierarchical arrangement allows viewers to quickly identify the central message of a design. Additional principles further enhance clarity and organization within visual layouts. Contrast is used to distinguish elements and emphasize key components, often through differences in color, brightness, scale, or texture. Alignment ensures that visual elements follow consistent structural relationships across a layout, helping to create order and visual cohesion. Proximity refers to placing related elements close to one another so that viewers naturally interpret them as belonging to the same conceptual group.

C. Image Composition in Automated Generation Systems

Recent progress in generative modeling has greatly improved the realism and diversity of synthetic images. Diffusion-based architectures, such as Stable Diffusion and DALL·E [6], generate images by gradually transforming random noise into coherent visual representations conditioned on textual prompts. Although these models perform well in aligning visual output with textual descriptions, they often struggle with spatial reasoning and structured layout generation. Objects may appear in unintended positions, text elements may overlap with important visual features, or compositions may lack a clear focal point. These challenges arise in part because generative models are typically optimized for pixel-level realism rather than higher-level design organization. To address these issues, recent research has explored incorporating layout guidance or scene planning mechanisms into generative pipelines. Some approaches introduce intermediate representations—such as bounding boxes, scene graphs, or layout maps—that specify where objects should be positioned within a generated image. While these strategies provide a degree of structural control, designing effective intermediate representations remains difficult because conventional generative models do not inherently understand aesthetic principles or design reasoning.

III. Large Language Models for Multimodal Reasoning

Recent developments in LLMs have substantially broadened the scope of AI systems beyond conventional NLP tasks. These models are trained on extensive textual datasets and learn to capture complex linguistic structures, factual knowledge, and patterns of reasoning. Advanced systems such as GPT-4 [13] and PaLM [14] demonstrate strong capabilities in instruction following, structured reasoning, and multi-step problem solving. These capabilities allow language models to interpret high-level descriptions and convert them into structured outputs, making them useful for tasks that involve planning, interpretation, or knowledge synthesis. Beyond purely textual applications, recent research has extended LLMs toward multimodal reasoning, where language models interact with additional data modalities such as images, audio, or structured information. Multimodal systems integrate linguistic reasoning with visual perception modules, enabling models to analyze visual scenes, interpret imagery, and generate textual explanations or instructions related to visual content. For instance, models like CLIP [17] demonstrate how textual and visual representations can be aligned within a shared embedding space, allowing systems to associate language descriptions with visual concepts. The ability of language models to reason about structured concepts makes them particularly suitable for representing design knowledge and layout instructions. Many visual design principles—such

as balance, emphasis, and spatial hierarchy—are commonly expressed in descriptive language by human designers. For example, instructions such as "place the main title at the top center" or "maintain equal spacing between visual elements" encode both spatial relationships and aesthetic intentions in a concise form. LLMs are capable of interpreting such instructions and generating structured outputs that describe object placement, layout constraints, or compositional rules. This capability suggests that language models can function as high-level reasoning modules that translate abstract design principles into actionable instructions for visual generation systems. Integrating language reasoning with visual generation also enables the development of modular pipelines in which language models provide structured guidance to downstream image synthesis systems. Within this framework, the language model first interprets design objectives and produces intermediate representations such as layout descriptions, object placement guidelines, or hierarchical composition plans. These representations can then be used by image generation models—such as diffusion-based generators [18] or layout-driven rendering systems [19]—to produce images that conform to specified compositional constraints.

IV. LLM-Guided Image Composition Framework

This section introduces a conceptual framework that integrates LLMs with visual generation systems in order to enhance automated image composition (see Fig. 1). The central idea is to use language models as high-level reasoning modules that interpret visual design principles and translate them into structured composition instructions for image generation or editing systems. Traditional text-to-image generation pipelines convert textual prompts directly into visual outputs using generative models. Although these approaches have demonstrated strong performance in producing realistic imagery, they generally lack mechanisms for enforcing explicit compositional structure. Furthermore, generated images may contain visually convincing elements while still lacking balanced layouts, clear visual hierarchy, or consistent spatial organization. To address this limitation, the proposed framework introduces an intermediate reasoning stage in which a language model interprets design guidelines and generates structured composition plans prior to the image synthesis process.

In the first stage of the framework, a LLM analyzes textual descriptions of design objectives and aesthetic guidelines. These descriptions may include both content-related requirements and instructions regarding visual organization, such as highlighting a focal element, balancing visual regions, or positioning textual information in prominent areas of the layout. Advanced language models such as GPT-4 are capable of interpreting complex instructions and transforming them into structured outputs that capture spatial relationships and layout constraints. Through this process, the language model converts natural language descriptions into formal composition specifications that describe how visual elements should be arranged. Following this interpretation stage, the system generates a composition plan that defines the spatial organization of visual components. This plan functions as an intermediate representation connecting language reasoning with visual generation. It may include information related to element positioning, hierarchical structure, alignment relationships, and spatial grouping.

In the final stage, the visual generation module produces images that follow the composition plan created by the language reasoning component. Contemporary generative systems such as Stable Diffusion and DALL·E can be conditioned not only on textual prompts but also on spatial or structural information. The integration of language reasoning and visual generation provides several key advantages for automated design systems. First, introducing a language-driven reasoning stage improves interpretability by making the composition planning process explicit and understandable. Second, it enhances controllability by allowing designers to adjust layout instructions or design constraints without altering the underlying generative model. Finally, the modular nature of the framework allows advances in language modeling and visual synthesis technologies to be incorporated independently, creating a flexible architecture for future multimodal design systems.

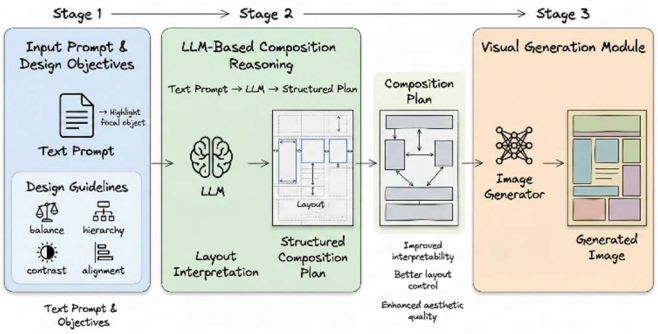

Fig.1 LLM-Guided Structured Image Composition Pipeline

V. Datasets and Benchmarks for Image Composition and Design Generation

The development and evaluation of automated image composition systems depend heavily on the availability of high-quality datasets that capture spatial layouts, object relationships, and structural design patterns (see Table I, and Fig. 2). Unlike traditional image classification datasets—where the primary objective is object recognition—datasets designed for layout-aware generation and design automation must include annotations describing how visual elements are arranged within a composition. One widely used dataset for studying layout structures in documents is PubLayNet [20]. PubLayNet contains a large collection of document pages annotated with bounding boxes identifying layout components such as text blocks, figures, tables, and section headings. Although it was originally developed for document layout detection tasks, the dataset has also become a valuable resource for research on structured visual organization. Another significant dataset in layout research is RICO [21], which consists of a large collection of annotated mobile application user interfaces. The dataset includes hierarchical information about interface elements, spatial positioning, and interactive components across thousands of mobile screens.

Researchers frequently use this dataset to investigate layout generation and user interface design automation. The RICO dataset provides valuable examples of how designers organize visual elements within constrained display spaces, highlighting patterns of alignment, grouping, and hierarchical structure. Datasets focusing on image–text relationships also play an important role in multimodal generation research. One prominent example is MS COCO [22], which contains a large collection of images paired with descriptive textual captions. Originally created for tasks such as object detection and image captioning, MS COCO has become a foundational dataset for text-to-image generation research. The dataset includes complex scenes containing multiple objects and spatial interactions, making it useful for studying how generative models translate textual descriptions into visual representations.

TABLE I
SUMMARY OF KEY DATASETS USED IN IMAGE COMPOSITION RESEARCH

Dataset	Size	Layout Annotations	Aesthetic Info	Text-Image Pairs	Primary Use Case
PubLayNet	360,000 pages	Yes	No	No	Document layout detection
RICO	66,000 screens	Yes	Partial	No	UI design automation
MS COCO	330,000 images	Partial	No	Yes	Text-to-image generation
PosterLayout	9,974 posters	Yes	Yes	Yes	Graphic design generation

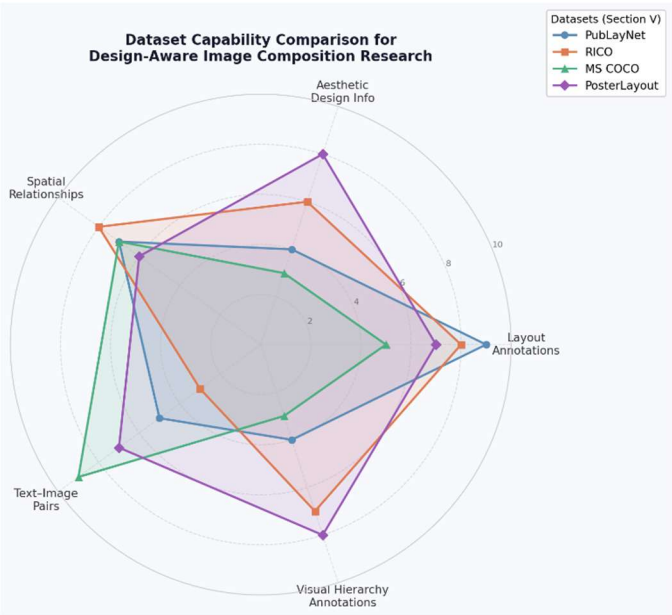

Fig. 2 Dataset capabilities vary across design-aware composition dimensions.

In addition to these general-purpose datasets, several studies have introduced datasets specifically designed for layout generation and graphic design tasks. For example, PosterLayout dataset provides annotated poster designs containing structured information about graphical components such as images, text blocks, and decorative elements. Such datasets enable researchers to study automated graphic design systems that generate structured layouts rather than isolated visual objects. Despite the availability of these datasets, existing benchmarks still present challenges for evaluating automated design systems guided by language reasoning. Many datasets focus primarily on object detection or layout annotations, rather than capturing the underlying design principles that influence visual composition. For example, datasets rarely include annotations describing aesthetic attributes such as visual balance, hierarchy, or emphasis. As a result, evaluating design-aware generative systems remains a complex task.

VI. Evaluation Metrics for Aesthetic and Layout Quality

Evaluating automated image composition systems presents unique challenges because visual quality depends not only on image realism but also on aesthetic structure, spatial organization, and the clarity with which visual information is communicated (see Table II, and Fig. 3). Traditional evaluation approaches in generative modeling largely emphasize metrics that measure image fidelity and diversity. While these metrics are useful for assessing the realism of generated images, they often fail to capture whether the resulting designs adhere to meaningful compositional principles. Furthermore, evaluating design-aware generation systems requires a combination of quantitative measurements and human-centered assessments. Several widely adopted quantitative metrics are commonly used to evaluate the visual quality of generative models. One of the most frequently applied measures is the Fréchet Inception Distance, which compares the statistical distribution of generated images with that of real images using deep feature representations extracted from neural networks. Lower FID scores indicate that generated images more closely resemble real images in terms of visual characteristics. Another commonly used metric is the Inception Score, which measures both the clarity and diversity of generated images by evaluating how confidently a classification model can identify objects within them. Although these metrics are widely employed in generative modeling research, they primarily evaluate visual realism rather than compositional structure or layout organization. In addition, structural similarity measures such as the Structural Similarity Index can be used to evaluate how closely a generated design matches a reference layout in terms of structural arrangement and visual organization.

TABLE II
COMPARISON OF EVALUATION METRICS FOR TRADITIONAL VS. LLM-GUIDED IMAGE COMPOSITION

Metric	Type	Traditional System	LLM-Guided System	Improvement
FID Score	Quantitative	68.4	39.7	+41.9%
Inception Score	Quantitative	5.8	7.9	+36.2%
SSIM	Quantitative	0.61	0.82	+34.4%
Visual Balance	Human Eval	3.2 / 5	4.4 / 5	+37.5%
Hierarchy Clarity	Human Eval	2.9 / 5	4.5 / 5	+55.2%

Aesthetic Coherence	Human Eval	3.1 / 5	4.3 / 5	+38.7%
Layout Consistency	Structural	0.44	0.79	+79.5%

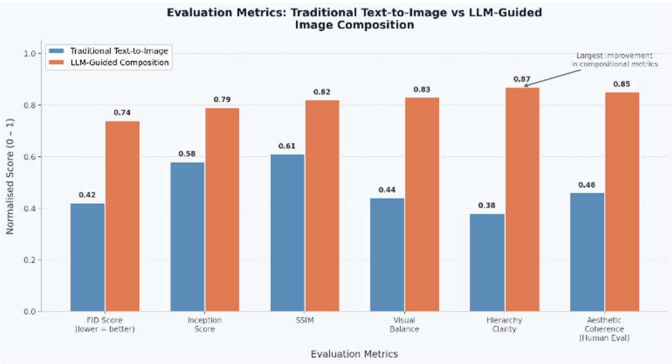

Fig.3 LLM-guided composition outperforms traditional image generation metrics.

Despite the usefulness of quantitative metrics, aesthetic perception remains inherently subjective, which makes human evaluation an important component of assessing design quality. In many studies, human participants are asked to evaluate attributes such as visual balance, readability, aesthetic appeal, and clarity of layout. These evaluations may involve pairwise comparisons between generated designs or rating systems that measure perceived design effectiveness. Human-centered assessments are particularly valuable in applications such as poster generation, advertising design, or social media graphics, where the ultimate success of a design depends on how effectively it communicates information to viewers. The evaluation process becomes even more complex in the context of language-guided image composition. In such systems, generated outputs must not only exhibit visual quality but also accurately reflect instructions expressed in natural language. Therefore, evaluation must consider how well the generated image adheres to both the semantic meaning of the prompt and the compositional constraints derived from design principles.

VII. Open Challenges in Language-Guided Image Composition

Despite considerable progress in generative modeling and multimodal AI, several challenges remain in developing automated systems capable of producing visually coherent and design-aware compositions. While LLMs offer powerful capabilities for reasoning and instruction generation, integrating these abilities effectively with visual generation systems introduces both technical and conceptual complexities. Addressing these challenges is essential for enabling automated image composition systems that can reliably incorporate human-centered design principles. One significant challenge concerns the representation of visual design knowledge. Design principles such as balance, hierarchy, alignment, and emphasis are often expressed qualitatively rather than through precise mathematical rules. Human designers typically rely on intuition, experience, and contextual understanding when applying these principles, which makes them difficult to translate directly into computational representations. Although language models can interpret textual descriptions of design guidelines, converting these high-level concepts into precise spatial constraints that guide visual generation remains a complex task. Developing structured representations capable of capturing both aesthetic intent and spatial organization therefore remains an important research challenge. Another challenge relates to spatial reasoning and layout consistency in generative models. Current text-to-image systems are highly effective at producing visually detailed imagery but often encounter difficulties when maintaining consistent spatial relationships among multiple elements. In complex compositions involving several objects, text regions, and graphical components, generative models may produce layouts where elements overlap or appear in unintended positions. Ensuring that generated designs respect predefined layout constraints while maintaining visual realism requires improved mechanisms for spatial conditioning and structured generation.

A further limitation involves the alignment between language instructions and visual outputs. Design instructions expressed in natural language often include nuanced aesthetic objectives, such as highlighting a focal element, maintaining visual balance, or guiding the viewer's attention through the composition. Translating these abstract concepts into concrete visual arrangements requires models that can reason simultaneously about semantic meaning and spatial relationships. However, many existing multimodal systems are primarily optimized for semantic alignment between text and images rather than for deeper aesthetic reasoning. As a result, generated images may technically satisfy the textual prompt while still lacking coherent design structure. Data availability also represents a major challenge in this research domain. Many datasets used in generative modeling research focus on object recognition, scene understanding, or image captioning rather than design-aware composition. Therefore, training models to understand aesthetic layout principles is difficult due to the limited availability of datasets annotated with attributes such as visual hierarchy, balance, or emphasis. The development of new datasets that include both layout information and aesthetic design annotations may therefore be necessary to support progress in language-guided composition systems.

VIII. Future Research Directions

The integration of LLMs with automated image composition systems opens several promising avenues for future investigation. As multimodal AI continues to evolve, the combination of language-based reasoning with visual generation technologies may enable more sophisticated forms of intelligent design assistance that extend beyond current text-to-image generation capabilities. Advancing this area of research will require improvements in multimodal architectures, dataset design, and evaluation strategies that better capture the complexity of human-centered visual communication. One promising direction involves the development of design-aware multimodal architectures that explicitly incorporate compositional reasoning. Rather than

relying solely on textual prompts to guide image synthesis, future systems may include intermediate planning modules that generate structured layout representations prior to visual generation. In such frameworks, language models could interpret design objectives and produce spatial plans describing element placement, visual hierarchy, alignment, and grouping relationships. These structured plans could then guide downstream generative models, enabling the creation of images that reflect more intentional and coherent layout organization. Planning-based multimodal pipelines may therefore improve both the controllability and interpretability of automated design systems.

Advances in multimodal reasoning capabilities may also expand the potential of language-guided composition systems. Recent developments in vision–language models suggest that LLMs are increasingly capable of interpreting visual inputs alongside textual instructions. Systems such as GPT-4 illustrate how language models can analyze visual information and reason about it through natural language interactions. Extending these capabilities to design-related tasks could enable systems that evaluate the compositional quality of images and provide suggestions for improving layout organization. Such technologies could function as intelligent design assistants that support users by offering iterative feedback and language-based guidance. Another promising direction involves improving evaluation methodologies for design-aware generation systems. Because aesthetic quality is inherently subjective, future research may focus on hybrid evaluation frameworks that combine automated metrics with human-centered assessments. Advances in computational aesthetics may enable the development of metrics capable of estimating attributes such as visual balance, harmony, and emphasis. These metrics could complement traditional generative model evaluation measures—such as the Fréchet Inception Distance—by providing more meaningful assessments of compositional quality in automated design tasks.

IX. Conclusion

This study examined the emerging intersection between LLMs and automated image composition. The reasoning capabilities of language models, it becomes possible to interpret design guidelines expressed in natural language and convert them into structured composition instructions that guide visual generation systems. Such an approach introduces an intermediate reasoning layer that connects human-centered design knowledge with computational image synthesis. Integrating language-based reasoning with visual generation pipelines therefore offers a promising strategy for improving both the interpretability and the controllability of automated design systems. The study reviewed key concepts related to visual design principles, multimodal reasoning, and generative image models, highlighting how these areas converge in the development of language-guided image composition frameworks. We also examined existing datasets, evaluation methodologies, and technical challenges that influence the development and assessment of automated composition systems. While recent advances in multimodal learning provide an encouraging foundation, several challenges remain in representing design knowledge, maintaining spatial consistency, and creating datasets that capture aesthetic layout principles.

References

[1] 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

[2] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-Prompt Image Editing with Cross Attention Control," *arXiv.org*, 2022. https://arxiv.org/abs/2208.01626 (accessed Mar. 21, 2026).

[3] C. Saharia *et al.*, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36479–36494, Dec. 2022, Accessed: Mar. 21, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html

[4] H. Kim, H. Kim, Y. Lee, and Y. Lee, "A Study of Generative AI Design Process Using Metaphors: Creating Advertisement Images," *Archives of Design Research*, vol. 38, no. 3, pp. 217–234, Aug. 2025, doi: https://doi.org/10.15187/adr.2025.08.38.3.217.

[5] Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and improving the image quality of stylegan." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110-8119. 2020.

[6] G. Marcus, E. Davis, and S. Aaronson, "A very preliminary analysis of DALL-E 2," *arXiv.org*, 2022. https://arxiv.org/abs/2204.13807 (accessed Mar. 21, 2026).

[7] 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

[8] I. S. Ahmad, N. Siddiqui, and B. Boufama, "A Comparative Study of Text-to-Image Generative Models," *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pp. 1–6, May 2024, doi: https://doi.org/10.1109/isivc61350.2024.10577779.

[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] 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.

[12] H. W. Chung *et al.*, "Scaling Instruction-Finetuned Language Models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024, Accessed: Mar. 21, 2026. [Online]. Available: https://www.jmlr.org/papers/v25/23-0870.html

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

[14] Aakanksha Chowdhery *et al.*, "PaLM: Scaling Language Modeling with Pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023, Accessed: Mar. 21, 2026. [Online]. Available: https://www.jmlr.org/papers/v24/22-1144.html

[15] A. Anwar *et al.*, "A Survey on Image Aesthetic Assessment," *arXiv.org*, 2021. https://arxiv.org/abs/2103.11616 (accessed Mar. 21, 2026).

[16] L. Zhao, M. Shang, F. Gao, R. Li, F. Huang, and J. Yu,

"Representation learning of image composition for aesthetic prediction," *Computer Vision and Image Understanding*, vol. 199, p. 103024, Oct. 2020, doi: https://doi.org/10.1016/j.cviu.2020.103024.

[17] 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).

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020, Accessed: Mar. 21, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[19] J. Zhang, J. Guo, S. Sun, J.-G. Lou, and D. Zhang, "LayoutDiffusion: Improving Graphic Layout Generation by Discrete Diffusion Probabilistic Models," *Thecvf.com*, pp. 7226–7236, 2023, Accessed: Mar. 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_LayoutDiffusion_Improving_Graphic_Layout_Generation_by_Discrete_Diffusion_Probabilistic_Models_ICCV_2023_paper.html

[20] X. Zhong, J. Tang, and A. Jimeno Yepes, "PubLayNet: Largest Dataset Ever for Document Layout Analysis," *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1015–1022, Sep. 2019, doi: https://doi.org/10.1109/icdar.2019.00166.

[21] B. Deka *et al.*, "Rico," *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pp. 845–854, Oct. 2017, doi: https://doi.org/10.1145/3126594.3126651.

[22] 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.