applied sciences

MDPI

Article

A Comparative Analysis of Generative AI for Architectural Visualization: A Quantitative Comparison Using Objective Evaluation Metrics and the Architectural AI Score (AIS)

Volkan Cantemir [1], Minel Kurtuluş [2] and Ece Cantemir [3,*]

[1] Cyber Security Program, Istanbul Topkapi University, Istanbul 34020, Türkiye; volkancantemir@topkapi.edu.tr
[2] Department of Interior Architecture, Istanbul Gelisim University, Istanbul 34310, Türkiye; mikurtulus@gelisim.edu.tr
[3] Department of Interior Architecture and Environmental Design, Istanbul Gelisim University, Istanbul 34310, Türkiye
* Correspondence: ecantemir@gelisim.edu.tr

Abstract

This study presents a quantitative metric-based comparison of four state-of-the-art text-to-image generative AI systems for architectural visualization: OpenAI Image Generation, Imagen 4, Midjourney v8.1, and Stable Image Ultra. The models were evaluated using the Architectural Prompt Dataset (APD-50), a controlled, expert-informed prompt set consisting of 50 architectural prompts across five typologies: Modern Residential Buildings, Public Buildings, Religious and Cultural Architecture, Interior Design, and Facade and Sustainable Architecture. For each prompt, four images were generated per model–prompt combination, resulting in a corpus of 800 images. The generated images were analyzed using reference-free and image-level computer vision metrics, including CLIP Score for prompt alignment, Raw BRISQUE and Quality_BRISQUE for reference-free perceptual quality, Shannon Entropy as a diagnostic descriptor of visual information density, LPIPS as intra-prompt perceptual variation, and SSIM as intra-prompt structural similarity. In response to the limitations of treating entropy as an inherently positive indicator, Shannon Entropy was excluded from the revised composite score and retained only as a diagnostic metric. Alongside the individual metric results, this study proposes the Revised Architectural AI Score (AIS-R) as a first-stage, exploratory composite index for summarizing selected visual-output characteristics under controlled prompt conditions. AIS-R combines normalized CLIP, Quality_BRISQUE, and LPIPS components using an explicit and reproducible weighting structure. The purpose of AIS-R is not to replace professional architectural judgment or to establish a fully validated architectural-quality framework at this stage. Rather, it provides an initial operational formulation for comparing prompt alignment, reference-free perceptual quality, and intra-prompt perceptual variation within the limited scope of the APD-50 experimental setup. Linear mixed-effects models with prompt-level random intercepts were used as the primary inferential framework. Under the revised metric configuration, Midjourney v8.1 achieved the highest AIS-R score (0.6504 ± 0.0586), followed by Stable Image Ultra (0.6267 ± 0.0582), Imagen 4 (0.6069 ± 0.0707), and OpenAI Image Generation (0.5829 ± 0.0555). However, these differences should be interpreted as metric-based differences in selected visual-output characteristics rather than as evidence of architectural quality, constructability, spatial validity, scale accuracy, or professional workflow suitability. The study therefore represents an initial metric-formulation and controlled benchmarking stage, while future research should examine the correspondence between AIS-R and professional architectural judgments through expert evaluation, user studies, or workflow-based validation.

Check for updates

Academic Editor: Paris Kitsos

Received: 17 July 2026
Revised: 5 September 2026
Accepted: 7 September 2026
Published: 6 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: generative AI; architectural visualization; diffusion models; revised architectural AI score; multi-metric benchmarking

1. Introduction

With advancements in information technologies, architectural visualization has evolved beyond traditional hand-drawing and sketching techniques, transforming through digital modeling, computer-aided design (CAD), and rendering technologies [1]. The integration of generative artificial intelligence technologies into architectural visualization has enabled designers to represent material qualities and environmental conditions with increasing accuracy and efficiency [2]. Software-based generative AI models are capable of producing original outputs through data-driven learning processes. This development has accelerated conceptual advancement, allowing abstract ideas and design intentions to be tested in an experimental setting during the early stages of architectural education and the design process [1,3].

Generative AI tools such as OpenAI Image Generation via the ChatGPT Pro web interface, Midjourney v8.1 via the Midjourney web interface, Stable Image Ultra via the BrandStudio web interface, and Imagen 4 via the Gemini web interface have emerged as prominent platforms for generating visual content from text-based prompts [3,4]. These platforms are capable of visualizing the material and atmospheric components of spaces with increasing realism [1]. However, the generated images vary in terms of architectural accuracy, visual style, usability, and suitability for different design stages [1,3].

In the literature, the treatment of artificial intelligence within design-oriented disciplines is generally limited to the use of individual tools or qualitative experimental applications [1,4]. There is a gap in the literature regarding comparative studies that examine how different text-to-image platforms perform in architectural visualization under controlled prompt conditions and multiple objective image-level metrics [5]. To address this gap, the present study compares four generative AI platforms—OpenAI Image Generation, Imagen 4, Midjourney v8.1, and Stable Image Ultra—using the APD-50 controlled prompt set. The dataset was created using prompts developed with expert input and represents five architectural visualization categories. Four independent images were generated per model for each prompt, resulting in a total corpus of 800 images. Rather than relying on subjective visual impressions alone, the study evaluates the generated outputs using multiple reference-free and image-level computer vision metrics, including prompt alignment, reference-free perceptual quality, diagnostic information density, intra-prompt perceptual variation, and intra-prompt structural similarity. However, these metrics are not treated as direct measures of architectural quality, constructability, spatial validity, scale accuracy, or professional design value. The study also proposes AIS-R as a first-stage exploratory composite index for summarizing selected visual-output characteristics under controlled prompt conditions. AIS-R is not intended to replace professional architectural judgment or to function as a fully validated architectural-quality framework at this stage. Instead, it provides an initial operational formulation for comparing selected image-level characteristics of AI-generated architectural visualizations in a transparent and reproducible way. Linear mixed-effects models, prompt-level aggregation, and exploratory Model × Category interaction analysis are then used to characterize model-level and category-level variation while accounting for prompt-level dependency. The resulting findings are therefore interpreted as controlled metric-based comparisons and as a foundation for future expert-based validation, rather than as definitive claims of architectural workflow suitability.

Research Questions

1. How do state-of-the-art text-to-image generative AI platforms differ across individual image-level metric profiles, including prompt alignment, reference-free perceptual quality, diagnostic information density, intra-prompt perceptual variation, and intra-prompt structural similarity, under the APD-50 controlled prompt set?
2. How can AIS-R be operationalized as a first-stage exploratory composite index for summarizing selected visual-output characteristics, and how sensitive are model rankings to alternative entropy-free weighting assumptions?
3. How do linear mixed-effects models, prompt-level aggregation, and exploratory Model × Category interaction analysis characterize model-level and category-level variation while accounting for prompt-level dependency?

2. Literature Review

2.1. Artificial Intelligence in Architecture

The use of artificial intelligence in architecture generally centers on three key areas. The first involves its role as a conceptual design accelerator, supporting the generation of alternative design solutions and early-stage idea development [1,5]. The second concerns its application as a tool for technical precision, specifically through integration with parametric modeling, façade optimization, building information modeling (BIM), and performance-oriented architectural simulations [6]. The third area involves the use of artificial intelligence for educational and experimental purposes [3]. However, most studies either examine individual AI tools or address artificial intelligence from a broad theoretical perspective. Comparative studies evaluating multiple state-of-the-art generative AI models across different stages of the architectural workflow remain limited [5]. Consequently, there is still a need for systematic analyses that can objectively compare these systems while taking into account the practical requirements of architectural visualization and conceptual design. Artificial intelligence is increasingly recognized as a tool that supports design rather than a technology that replaces architects [1,4].

Artificial intelligence systems leveraging machine learning and design automation can process extensive architectural datasets, generate alternative building forms, optimize structural configurations, and assist designers during conceptual development [5,7]. In particular, evolutionary algorithms and deep learning techniques enrich the design process by enabling architects to generate diverse design alternatives [5,6]. This interdisciplinary approach shifts the architect's role from manually producing visualizations to overseeing and guiding generative workflows, thereby positioning artificial intelligence as a supportive tool for the designer [2,6]. Furthermore, the integration of big data, cloud computing, and distributed artificial intelligence into digital platforms such as BIM and CAD has significantly enhanced architectural data analysis and complex environmental simulations [2,6]. These advancements enable architects to make more informed design decisions while increasing computational efficiency throughout the design process [2,5]. Alongside these advantages, issues regarding cost, usability, and ethical practices are also documented in the literature, indicating that the impact of artificial intelligence on the architectural discipline remains a subject of critical debate [2,7].

Recent studies demonstrate that integrating generative AI with computational and parametric design environments accelerates the conceptual design process, supports the generation of alternative forms, and contributes significantly to decision-making during the early stages of architectural design [6]. Furthermore, AI-generated visualizations enhance communication among architects, clients, and other stakeholders during the conceptual design phase by transforming parametric concepts into highly realistic renderings [1,5].

Additionally, AI-based decision support and rapid prototyping systems enable the rapid evaluation of design alternatives and their effective refinement based on user feedback [1,5].

2.2. Human–Artificial Intelligence Collaboration

In the application of artificial intelligence to architecture, the active involvement of designers in aspects such as contextual understanding, functional requirements, aesthetic evaluation, and spatial reasoning ensures that AI is positioned not as a system replacing the architect, but rather as an assistive tool that accelerates the design process and fosters creativity [6,8]. The significant contributions AI-driven systems make to idea generation, the creation of design alternatives, and data analysis facilitate the development of architectural designs within an experimental framework [1,3]. However, it is evident that the integration of these technologies into architectural workflows has not yet been fully standardized [2]. This indicates that the effective use of AI in architecture depends not only on technological advancements, but also on user experience, design methodologies, and organizational adoption processes [1,2]. Nevertheless, AI-supported design processes also entail certain significant challenges. Researchers highlight risks related to copyright, authorship, algorithmic bias, and ethical responsibilities, alongside technical issues such as usability, computational costs, and system adoption [1].

2.3. Prompt Engineering

Prompt engineering is being employed as a new design methodology in architecture, just as it is in many other disciplines. Unlike traditional software interfaces, text-to-image systems utilize textual prompts to convey design intentions [3,4]. Architectural prompts typically include detailed descriptions of building typology, structural systems, façade materials, lighting conditions, architectural style, environmental context, and rendering quality. Phrases such as “minimalist concrete villa”, “double-skin glass façade”, “parametric pavilion”, or “biophilic interior” enable designers to control visual appearance alongside semantic content [4]. Furthermore, recent studies indicate that output quality depends significantly not only on prompt content, but also on user experience, prompt-writing strategies, and model-specific parameter settings [3,4]. In particular, auxiliary models such as ControlNet and LoRA—used within the Stable Diffusion ecosystem—allow users to achieve a higher level of control over composition, perspective, and structural accuracy [8].

2.4. Diffusion Models for Architectural Design

Diffusion models have fundamentally transformed architectural visualization by generating highly realistic images through processes that iteratively remove noise. Compared to earlier generative adversarial networks (GANs), diffusion-based approaches offer superior semantic consistency, improved texture generation, and higher prompt fidelity [8]. State-of-the-art systems such as Stable Image Ultra, Midjourney, OpenAI Image Generation, and Google’s Imagen have widely been adopted for conceptual architectural visualization [2,4]. Stable Diffusion enables architects to exercise greater control over composition and structural consistency by offering extensive parameter customization and integration with extensions like ControlNet and LoRA [2,8]. While Midjourney has gained popularity for its artistic rendering quality and atmospheric visualizations, OpenAI Image Generation and Imagen demonstrate strong semantic understanding and realistic image synthesis [2,3]. Recent studies indicate that diffusion models are successfully employed not only in conceptual visualization, but also in a wide range of architectural applications, including facade design, digital reconstruction of historical heritage, interior design, sustainable architecture, and conceptual urban design [2,8].

2.5. Architectural Visualization

Architectural visualization is important for realistically perceiving qualities such as spatial organization, material authenticity, and lighting within a project. With the emergence of generative AI, the visualization process has accelerated significantly through the transformation of conceptual ideas, sketches, or simplified massing models into renderings [8]. Although these technologies substantially reduce visualization time, previous research indicates that human oversight remains indispensable for ensuring contextual consistency, architectural feasibility, and design quality [1].

Another application area of architectural visualization is architectural education [2]. Recent studies indicate that generative AI enhances creativity, accelerates idea development, and improves student engagement in architectural design studios. Research shows that while some image-generating models may produce repetitive visual patterns in response to similar prompts [9], newer versions like Stable Diffusion offer greater diversity [8]. AI-supported educational environments enable students to rapidly visualize conceptual ideas, explore multiple design alternatives, and experiment with diverse architectural styles. Studies conducted in interior design education reveal that generative AI accelerates experimental design processes and increases student satisfaction [3]. In the context of biophilic design, systems such as Leonardo AI have been reported to support the development of nature-inspired spatial concepts [4].

2.6. Objective Evaluation of AI-Generated Architectural Images

Despite rapid advancements in image generation, the evaluation of AI-generated architectural visualizations remains a significant research challenge. Traditional computer vision metrics, such as Fréchet Inception Distance (FID) and Inception Score (IS), were originally developed for natural image datasets and exhibit significant limitations when applied to architectural visualization. FID requires extensive collections of reference images, which are often unavailable for imaginative architectural concepts. Similarly, Inception Score relies heavily on ImageNet object categories and fails to capture architectural semantics, spatial composition, façade organization, and building typologies. For these reasons, and as highlighted by Mittal et al. (2012) and Hessel et al. (2021), recent studies increasingly employ reference-free objective evaluation metrics in scenarios where reference images are unavailable [10,11]. In this context, the CLIP Score, introduced by Hessel et al. (2021) [11] as a reference free metric, measures the semantic alignment between textual prompts and generated images. Meanwhile, BRISQUE, developed by Mittal et al. (2012) [10], assesses perceptual image quality without requiring reference images. While Shannon Entropy describes pixel-intensity information density [12], it should not be interpreted as a direct indicator of architectural richness or design quality. Similarly, LPIPS, developed by Zhang et al. (2018) and used in generative image evaluation contexts by Zhu et al. (2017), measures feature-level perceptual variation among generated outputs but does not directly validate meaningful architectural design diversity or spatial innovation [9,13]. SSIM, introduced by Wang et al. (2004), reports local structural similarity between image pairs and should not be equated with architectural or engineering stability [14]. Together, these metrics provide reproducible descriptors of selected visual-output characteristics, but they do not directly assess architectural scale accuracy, constructability, spatial logic, or professional design value.

3. Methods

3.1. Dataset (APD-50)

In this study, the APD-50 (Architectural Prompt Dataset) was created based on the insights and academic feedback of experts holding degrees in architecture. To ensure

interdisciplinary inclusivity and typological diversity, five core visualization categories were selected: Modern Residential Buildings, Public Buildings, Religious and Cultural Architecture, Interior Design, and Facade and Sustainable Architecture. Ten standard prompts were developed for each category, resulting in a total of 50 prompts. To ensure comparability regarding output quality and rendering objectives, the phrase "photorealistic architectural visualization, highly detailed, cinematic lighting, realistic textures, 8K resolution" was appended as a common suffix to every prompt. This common suffix was deliberately chosen as a methodological standardization strategy to minimize stylistic drift between models, stabilize the target visual output, and maintain the comparison within a consistent photorealistic architectural visualization framework. It is acknowledged as a limitation that this decision may restrict the generalizability of the models' performance regarding other prompt writing styles.

The APD-50 prompt set was developed through a qualitative expert-informed process involving 12 academics working in the field of architecture. The expert group included assistant professors, lecturers, and research assistants with academic competence in architecture and related design fields. Initial prompts were drafted under five predefined visualization categories: Modern Residential Buildings, Public Buildings, Religious and Cultural Architecture, Interior Design, and Facade and Sustainable Architecture. Expert feedback was used to refine the prompts in terms of typological relevance, clarity of architectural terminology, suitability for text-to-image generation, and consistency with the intended photorealistic architectural visualization objective. No formal scoring, ranking, or consensus-rating procedure was conducted; therefore, no inter-expert agreement coefficient was calculated. Accordingly, APD-50 should be understood as a controlled, expert-informed prompt set rather than an expert-validated or universally representative benchmark. The complete list of APD-50 prompts is provided in Appendix A. The phrases "photorealistic architectural visualization, highly detailed, cinematic lighting, realistic textures, 8K resolution" were appended to all prompts listed in this table. The resulting representative outputs are shown in Figure 1.

It should be noted that APD-50 is not intended to represent a universal or exhaustive benchmark of architectural visualization. Rather, it is a controlled, expert-informed prompt set designed to enable comparative testing under standardized conditions. The categories and prompts were developed to cover a range of common architectural visualization themes; however, the resulting model rankings may depend on the specific composition, typological balance, and wording of these 50 prompts. This limitation is explicitly acknowledged in the limitations section.

3.2. Evaluated Models

Regarding the reproducibility and methodological transparency of four state-of-the-art image generation models representing diverse architectural use cases and underlying technologies, the inference parameters and system details are as follows (Table 1): OpenAI Image Generation images were produced via the ChatGPT Pro web interface without specific model selection; system parameters cannot be configured by the user within the ChatGPT Pro interface; Imagen 4 images were generated using the Google Gemini Advanced web interface with default parameters; Midjourney v8.1 images were produced via the official web interface using default stylization parameters; and Stable Image Ultra images were obtained through the BrandStudio web interface using default parameters. Inference operations were completed in June 2026.



Figure 1. Representative outputs from the four evaluated generative AI platforms across the five APD-50 architectural categories.

Table 1. Inference parameters and system details of the evaluated generative AI models.

Model	Provider	Access Method	Version/ Release	Inference Environment	Resolution/ Aspect Ratio	Parameter Settings	Output per Prompt
OpenAI Image Generation	OpenAI	ChatGPT Pro web interface	Not publicly disclosed by OpenAI.	ChatGPT Pro, web interface (June 2026)	1536 × 1024/approx. 3:2	Default system parameters	4
Midjourney v8.1	Midjourney Inc.	Midjourney web interface	v8.1 (HD)	Midjourney Web Interface (June 2026)	2048 × 2048/1:1	Standard style, HD quality, stylization (default)	4
Stable Image Ultra	Stability AI	BrandStudio Web interface	Stable Image Ultra	BrandStudio Web interface (June 2026)	1024 × 1024/1:1	Default settings	4
Imagen 4 (Gemini)	Google	Gemini Web Interface	Imagen 4	Gemini Advanced Interface (June 2026)	2816 × 1536/approx. 16:9	Default settings	4

Preprocessing standardization: Although the raw images generated by the models varied in resolution and aspect ratio, they were not compared in their raw state; instead, all images underwent a common preprocessing protocol prior to metric calculation. Consequently, the direct impact of raw resolution differences on the metrics was controlled. However, as the compositional effects of aspect ratio and resizing could not be entirely eliminated, this factor has been acknowledged as a methodological limitation of the study.

3.3. Quantitative Metrics

The selected metrics were used to characterize image-level visual-output properties rather than to directly evaluate architectural design quality. CLIP, BRISQUE, Shannon Entropy, LPIPS, and SSIM provide information about prompt alignment, reference-free perceptual quality, pixel-level information density, intra-prompt perceptual variation, and intra-prompt structural similarity. However, these metrics do not directly assess architectural scale accuracy, constructability, spatial functional logic, material correctness, structural rationality, or professional design value.

CLIP Score (semantic alignment): The CLIP Score measures the semantic alignment between the visual characteristics of the generated image and the textual description of the prompt. The original CLIPScore (CLIP-S) formula, developed by Hessel et al., calculates the cosine similarity between the textual caption c and the visual image v as follows:

$$\text{CLIP-S}(c,v) = w \times \max(\cos(c,v),0)$$

Here, the coefficient w is defined as a regression-based weight (e.g., w = 2.5) intended to optimize correlation with human assessment. The formula adapted in the study is:

$$\text{CLIP_Score}(I,T) = \max(0,\cos(v_I,v_T) \times 100)$$

Here, v_I and v_T are embedding vectors obtained from the pre-trained openai/clip-vit-base-patch32 image and text encoders, respectively. The coefficient w in the original formula was set to 100, in accordance with the general consensus in the literature comparing computer vision and generative models. The purpose of this adjustment is to map the semantic alignment similarity to a percentage scale within the [0, 100] range, enhance interpretability, and ensure normalization consistency when integrating the metrics into the revised AIS-R composite score.

Shannon Entropy (diagnostic information density): Shannon Entropy measures the information density based on the distribution of pixel intensities within an image. This value can serve as an indirect indicator of the density of visual textures and detail patterns; however, on its own, it does not necessarily signify superior architectural detailing, appropriate material usage, or higher-quality design. The original entropy formula, defined by Claude Shannon within the framework of information theory, measures the information content of a message source characterized by a discrete probability distribution:

$$H = -\sum_{x \in X} p(x) log_2 p(x)$$

In the formula adapted for the study, the image is converted to grayscale (L = 256 gray levels), and the entropy H is calculated from the normalized histogram probability distribution (p_i) of the pixel intensities:

$$H = -\sum_{i=0}^{L-1} p_i \, log_2(p_i)$$

Here, p_i represents the value ni/n (the ratio of the number of pixels at a given gray level to the total number of pixels) [15]. Shannon's original communication theory formula

has been adapted to the two-dimensional spatial plane in accordance with the digital image processing literature. This enables the measurement of pixel-intensity information density associated with visual texture, shading, and local image detail, without implying architectural quality or design superiority.

Shannon Entropy measures the statistical information density associated with the distribution of pixel intensities within an image. In this study, entropy was retained only as a diagnostic descriptor of visual information density and was not interpreted as a direct indicator of architectural richness, design quality, or rendering success. Higher entropy may correspond to dense texture or visual complexity in some images, but it may also reflect visual noise or excessive detail. Conversely, minimalist architectural styles, smooth concrete surfaces, curtain-wall facades, and visually restrained compositions may naturally produce lower entropy values without being inferior in architectural or visual quality. For this reason, Shannon Entropy was excluded from the revised AIS-R composite score.

BRISQUE (no-reference visual quality): BRISQUE scores image quality without a reference image based on deviations from natural scene statistics (NSS). Developed by Mittal et al., the BRISQUE model scores spatial-domain image distortions on a scale of [0, 100]:

$$\text{BRISQUE} \in [0, 100]$$

Here, 0 represents the best visual quality (zero distortion), while 100 represents the worst visual quality (maximum distortion). The adapted formula used in the article is

$$\text{Quality_BRISQUE} = 100 - \text{BRISQUE}$$

Since the original BRISQUE output functions as a "distortion/error metric" (where lower values are better), the score was adjusted to align with other positive-oriented metrics (where higher values are better). A linear inversion (100 − BRISQUE) was applied to ensure directional alignment with multi-criteria decision-making (MCDA) standards. Consequently, scores close to 100 indicate high visual quality and the absence of artifacts such as blurring or noise. Prior to calculation, images were standardized to 512 × 512 pixels to prevent resolution-related discrepancies.

LPIPS Perceptual Variation (intra-prompt perceptual distance): LPIPS was used to quantify intra-prompt perceptual variation among the four images generated for the same model–prompt combination. Zhang et al. formulated the perceptual distance (d_LPIPS) between the deep network activations of two image patches (x, x_0) as follows:

$$d_{LPIPS}(x, x_0) = \sum_l \frac{1}{H_1 W_1} \sum_{h,w} \| w_l \odot (\hat{y}\frac{l}{hw} - \hat{y}\frac{l,0}{hw}) \| \frac{2}{2}$$

This metric measures feature-level perceptual differences between image outputs; it does not directly measure meaningful architectural design diversity, spatial innovation, or functional design value. In this study, the average $LPIPS$ distance between all possible image pairs (C(4,2) = 6 pairs) was calculated for each model–prompt combination to describe intra-prompt perceptual variation among four generated outputs [9].

$$LPIPS_diversity = \frac{1}{6} \sum_{j<k} d_{LPIPS}\left(img_j, img_k\right)$$

Here, the images were resized to 256 × 256 pixels and normalized before being evaluated using the AlexNet-based LPIPS implementation. Because the original LPIPS formulation measures the distance between two images, this study adapts it by averaging pairwise distances across the four outputs generated for each model–prompt combination. Following the pairwise comparison logic used in generative image evaluation studies such

as Zhu et al. (2017) [9], the resulting value is interpreted only as a descriptor of perceptual variation, and not as evidence of higher architectural design quality or more valuable design alternatives.

SSIM Intra-Prompt Structural Similarity: $SSIM$ was used as an intra-prompt structural similarity descriptor by calculating the average pairwise similarity among images generated under the same model–prompt combination. In this context, $SSIM$ does not measure structural stability in the architectural or engineering sense. Rather, it indicates the degree of local luminance, contrast, and structural similarity between image outputs. Wang et al. defined the structural similarity between two images (x, y) by combining luminance, contrast, and structural components as follows:

$$SSIM(x,y) = \frac{\left(2\mu_x\,\mu_{\,y} + C_1\right)\left(2\sigma_{\,xy} + C_2\,\right)}{\left(\mu_x^2 \;+\; \mu_y^2 \;+\; C_1\right)\left(\sigma_x^2 \;+ \sigma_y^2 \;+\; C_2\,\right)}$$

In the adapted formulation used in this study, the average pairwise $SSIM$ was calculated across all possible image pairs for each model–prompt combination to describe intra-prompt output similarity.

$$\text{SSIM_consistency} = \frac{1}{6}\sum_{j<k} SSIM\left(img_j,\, img_k\right)$$

The original $SSIM$ compares a reference image with a degraded image. In this study, however, $SSIM$ was adapted to a no-reference multi-output setting by averaging pairwise $SSIM$ values among the four outputs generated for the same model–prompt combination. Similar to LPIPS, $SSIM$ was interpreted only as a supplementary descriptor of intra-prompt output similarity, not as a measure of architectural structural stability, geometric correctness, or design quality.

3.4. Revised Architectural AI Score (AIS-R)

In the original formulation, Shannon Entropy was included as a positively weighted component of the composite score. However, as entropy does not have a universally positive interpretation in architectural visualization, especially for minimalist or visually restrained design styles, it was removed from the revised composite formulation. The Revised Architectural AI Score (AIS-R) is therefore defined as a study-specific composite index for summarizing selected visual-output characteristics under controlled prompt conditions. It combines normalized CLIP Score, normalized Quality_BRISQUE, and normalized LPIPS Perceptual Variation:

$$\text{AIS-R} = 0.40 \times \text{CLIP_norm} + 0.35 \times \text{Aesthetic_norm} + 0.25 \times \text{LPIPS_norm}$$

where Aesthetic_norm is the Min–Max normalized version of Quality_BRISQUE, calculated as:

$$\text{Quality_BRISQUE} = 100 - \text{Raw_BRISQUE}$$

All components were Min–Max normalized across the 800-image dataset prior to aggregation. Shannon Entropy was excluded from AIS-R and retained only as a diagnostic descriptor of visual information density. SSIM was not included in the composite score and was retained as a supplementary descriptor of intra-prompt structural similarity.

The purpose of AIS-R is not to replace professional architectural judgment or to establish a fully validated architectural-quality framework at this stage. Rather, AIS-R represents a first-stage operationalization of a transparent composite metric for comparing selected visual-output characteristics in AI-generated architectural visualizations. In this sense,

the present study should be understood as the initial metric formulation and controlled benchmarking stage, while external validation against professional architectural judgment remains a necessary subsequent stage of research.

The weighting hierarchy of AIS-R was defined as an explicit and reproducible starting formulation rather than as an empirically validated weighting model. CLIP Score was assigned the highest weight because prompt–image correspondence is a necessary baseline condition in text-to-image architectural visualization: an image that does not respond to the architectural prompt cannot be considered successful within the scope of this controlled benchmark. Quality_BRISQUE was assigned the second-highest weight because reference-free perceptual quality is important for evaluating the visual usability of generated images, including the presence or absence of distortions, artifacts, or perceptual degradation. LPIPS was retained with a positive but lower weight because AIS-R is intended to summarize not only prompt alignment and reference-free perceptual quality, but also the degree of intra-prompt perceptual variation produced under identical textual conditions.

The positive contribution of LPIPS should not be interpreted as evidence of higher architectural design quality, better spatial organization, or more valuable design alternatives. Rather, within the limited scope of this study, it reflects the ability of a model to generate perceptually varied outputs from the same prompt. This dimension is relevant to the comparative characterization of generative behavior, particularly in exploratory visualization contexts, but it does not constitute a direct measure of architectural creativity, diversity, or design value. For this reason, LPIPS was assigned a lower weight than CLIP and Quality_BRISQUE, and its influence was further examined through entropy-free sensitivity scenarios.

AIS-R is therefore not a validated measure of architectural quality, constructability, spatial validity, scale accuracy, or professional design value; it is a study-specific and exploratory composite summary of selected visual-output characteristics. The weighting scheme is heuristic and should not be interpreted as universally valid. Future work should test the correspondence between AIS-R and professional architectural judgments through expert evaluation, Delphi procedures, AHP-based expert weighting, structured user studies, or regression-based calibration against architect-rated image assessments.

3.5. Linear Mixed-Effects Statistical Modeling

Because the dataset contains repeated generations from the same prompts, ordinary ANOVA was not used as the primary inferential framework. Instead, linear mixed-effects models (MixedLMs) were used to account for prompt-level dependency. The main model included model identity and architectural category as fixed effects, with prompt code specified as a random intercept: AIS-R ~ C(Model) + C(Category) + (1 | Prompt) The model was fitted using maximum likelihood estimation. The intercept corresponds to Imagen 4 in Category A, while model coefficients represent differences relative to Imagen 4 after controlling for architectural category. In addition to the main model, two robustness analyses were performed. First, a prompt × model aggregated model was fitted by averaging AIS-R across the four image generations for each prompt–model combination, resulting in 200 aggregated observations. Second, an exploratory Model × Category interaction model was fitted to examine whether model effects varied across architectural categories: AIS-R ~ C(Model) × C(Category) + (1 | Prompt). The interaction model was treated as exploratory and was not used to make definitive category-specific or workflow-level claims.

4. Experimental Results and Analysis

4.1. Global Performance Comparison

According to the descriptive statistics, Midjourney v8.1 achieved the highest average score in the baseline AIS-R scenario. The overall quantitative performance of the four models across 800 images is summarized in Table 2. Based on the findings, Midjourney v8.1 demonstrated the highest overall average (AIS-R = 0.6504). Stable Image Ultra (AIS-R = 0.6267) ranked second and Imagen 4 (AIS-R = 0.6069) third, while OpenAI Image Generation (AIS-R = 0.5829) placed last.

Table 2. Global performance results of generative AI models under the revised AIS-R configuration.

Model	CLIP Score	Raw BRISQUE ↓	Quality BRISQUE ↑	Shannon Entropy Diagnostic	LPIPS Perceptual Variation	SSIM Intra-Prompt Structural Similarity	AIS-R ↑
OpenAI Image Generation	31.29 ± 1.88	26.15 ± 5.82	73.85 ± 5.82	7.63 ± 0.21	0.483 ± 0.055	0.134 ± 0.053	0.5829 ± 0.0555
Imagen 4	30.42 ± 2.27	27.94 ± 10.69	72.06 ± 10.69	7.74 ± 0.14	0.571 ± 0.045	0.144 ± 0.058	0.6069 ± 0.0707
Midjourney (v8.1)	31.60 ± 1.99	28.79 ± 9.06	71.21 ± 9.06	7.70 ± 0.12	0.595 ± 0.046	0.103 ± 0.047	0.6504 ± 0.0586
Stable Image Ultra	32.07 ± 2.01	30.46 ± 14.45	69.54 ± 14.45	7.65 ± 0.19	0.539 ± 0.049	0.180 ± 0.083	0.6267 ± 0.0582

Note: ↑ indicates that higher values represent better performance, whereas ↓ indicates that lower values represent better performance.

Although Midjourney v8.1 achieved the highest AIS-R score under the revised metric configuration, AIS-R should not be interpreted as a validated measure of architectural quality. The values summarize selected visual-output characteristics under the APD-50 prompt set and the specified preprocessing and weighting scheme.

The radar chart (Figure 2) illustrates comparative profiles across prompt alignment, Quality_BRISQUE, LPIPS perceptual variation, and AIS-R. The figure does not include Shannon Entropy because entropy is retained only as a diagnostic descriptor and is not part of AIS-R.

4.2. Category-Based Performance Analysis

To examine the performance of the models in specific architectural domains, the category-based average AIS-R values are summarized in Table 3. The lower average AIS-R scores observed in Category B suggest that public buildings may represent a more challenging visualization domain for the models within this prompt set. In contrast to public buildings, the highest average performance levels were achieved in Category C (Religious and Cultural Architecture).

Table 3. Average AIS-R scores of models by architectural typology.

Category	OpenAI Image Generation	Imagen 4	Midjourney	Stable Image Ultra
A. Modern Residential Buildings	0.5706	0.5997	0.6430	0.6215
B. Public Buildings	0.5692	0.5650	0.6174	0.5950
C. Religious and Cultural Architecture	0.6170	0.6692	0.6957	0.6806
D. Interior Design	0.5914	0.6105	0.6483	0.6417
E. Facade and Sustainable Architecture	0.5661	0.5899	0.6475	0.5947

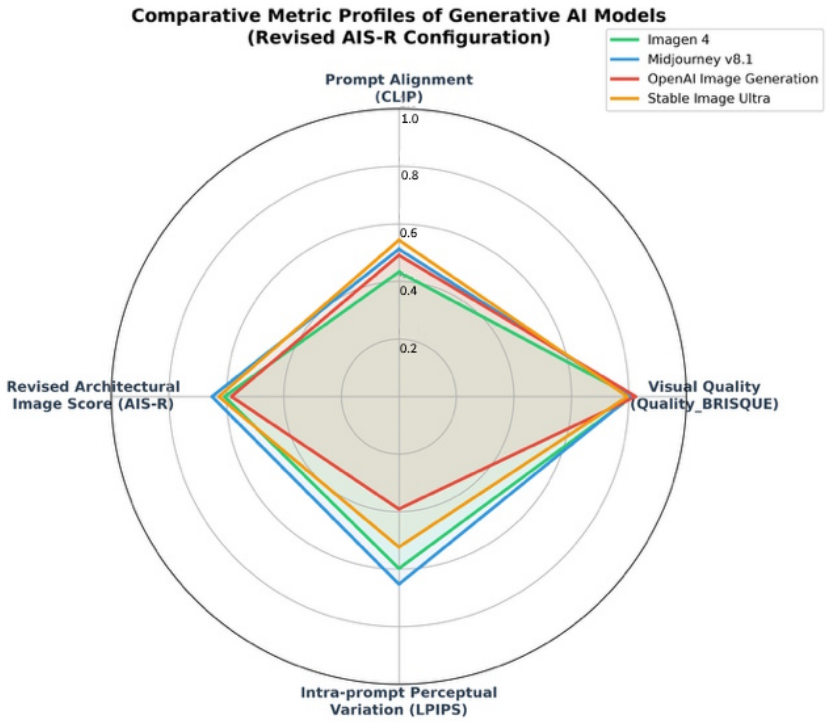


Figure 2. Comparative performance profiles (radar chart).

The category-wise values indicate that model performance profiles vary across prompt typologies. However, these category-level differences should be interpreted cautiously because the Model × Category interaction analysis was exploratory and the APD-50 categories do not constitute a universal representation of architectural visualization tasks.

The category-based heatmap (Figure 3) indicates that Category C (Religious and Cultural Architecture) produced relatively higher AIS-R values within the APD-50 prompt set.

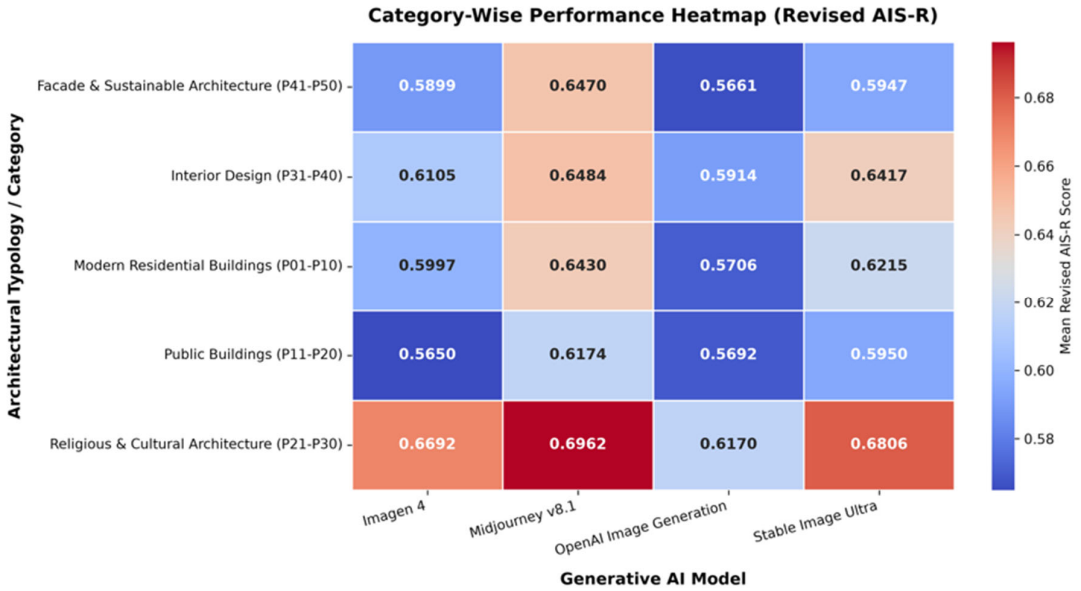


Figure 3. Category-based average AIS-R performance profile heat map.

4.3. Statistical Analysis: Mixed-Effects Regression Model

The main MixedLM model was fitted using ML estimation with AIS-R as the dependent variable, model identity and architectural category as fixed effects, and prompt code as a random intercept. The intercept corresponds to Imagen 4 in Category A, while model

coefficients represent differences relative to Imagen 4 after controlling for architectural category. Controlling for prompt-level random variance, Midjourney v8.1 showed a statistically significant positive coefficient relative to Imagen 4 ($\beta = +0.0435$, $z = 9.048$, $p < 0.001$, 95% CI [0.0341, 0.0530]). Stable Image Ultra also showed a smaller positive coefficient ($\beta = +0.0198$, $z = 4.120$, $p < 0.001$, 95% CI [0.0104, 0.0293]). OpenAI Image Generation showed a negative coefficient relative to Imagen 4 ($\beta = -0.0240$, $z = -4.987$, $p < 0.001$, 95% CI [−0.0334, −0.0146]). Model fit statistics were Log-Likelihood = 1249.49, AIC = −2478.99, and BIC = −2432.14.

4.4. Prompt × Model Aggregated Robustness Analysis

To address the concern that statistical significance may be driven by repeated image-level observations, AIS-R scores were aggregated at the prompt × model level. This produced 200 observations, corresponding to 50 prompts × 4 models. The aggregated robustness model yielded the same fixed-effect coefficient pattern as the image-level model, with the model main effects remaining statistically significant ($p \leq 0.008$). This supports that the observed model effects are not solely driven by repeated image-level sampling, although the practical magnitude of the differences remains modest.

4.5. Exploratory Model × Category Interaction Analysis

An exploratory Model × Category interaction model was fitted to examine whether model effects varied across architectural categories. The interaction model suggested some category-dependent variation (LR = 25.41, df = 12, $p = 0.0130$). However, because the interaction analysis was exploratory and model selection criteria were mixed, with AIC favoring the interaction model and BIC favoring the main model, category-specific interpretations were treated cautiously. The interaction results were not used to make definitive workflow-level claims.

4.6. Statistical Versus Practical Architectural Significance

Although several model coefficients reached statistical significance, the absolute differences in AIS-R were modest in practical magnitude. The observed differences between platforms were approximately ΔAIS-R $\approx$ 0.02–0.07 on a 0–1 scale. Therefore, statistically significant differences should not be interpreted as large or definitive architectural performance gaps. In particular, AIS-R differences do not imply that one platform produces architecturally superior, more constructible, or more professionally valid design outputs. Practical architectural significance requires future validation through expert evaluation, user studies, or real design workflow experiments.

4.7. Robustness and Weight Sensitivity Analysis

To examine the influence of heuristic weighting assumptions, five entropy-free weighting scenarios were tested: S0 Revised AIS-R, S1 Equal Weight, S2 Semantic-Heavy, S3 Aesthetic-Heavy, and S4 Variation-Heavy. Rankings were stable across S0–S3, while the variation-heavy scenario showed partial sensitivity, confirming that model rankings depend on weighting assumptions (Table 4). In S4, where LPIPS was assigned a higher weight, Imagen 4 slightly surpassed Stable Image Ultra (0.6126 vs. 0.6110). These results reinforce that AIS-R should be interpreted as a study-specific metric summary rather than a universally validated ranking system. This sensitivity analysis was not intended to establish the external validity of AIS-R or to prove that the selected weights represent professional architectural judgment. Rather, it was used as a transparency check to show how model rankings respond to alternative entropy-free weighting assumptions. Therefore, the results should be interpreted as evidence of weight-dependence and relative ranking

stability within the present dataset, not as validation of AIS-R as a universal architectural evaluation framework.

Table 4. Robustness and weight sensitivity analysis under entropy-free AIS-R scenarios.

Model	S0 Revised AIS-R Mean ± SD	S0 Rank	S1 Equal Weight Mean ± SD	S1 Rank	S2 Semantic-Heavy Mean ± SD	S2 Rank	S3 Aesthetic-Heavy Mean ± SD	S3 Rank	S4 Variation-Heavy Mean ± SD	S4 Rank
Midjourney v8.1	0.6504 ± 0.0586	1	0.6571 ± 0.0547	1	0.6287 ± 0.0680	1	0.6875 ± 0.0527	1	0.6566 ± 0.0549	1
Stable Image Ultra	0.6267 ± 0.0582	2	0.6208 ± 0.0530	2	0.6152 ± 0.0679	2	0.6654 ± 0.0569	2	0.6110 ± 0.0525	3
Imagen 4	0.6069 ± 0.0707	3	0.6143 ± 0.0648	3	0.5797 ± 0.0822	3	0.6555 ± 0.0624	3	0.6126 ± 0.0636	2
OpenAI Image Generation	0.5829 ± 0.0555	4	0.5689 ± 0.0544	4	0.5712 ± 0.0640	4	0.6378 ± 0.0466	4	0.5511 ± 0.0574	4

Note: S0 = Revised AIS-R baseline (CLIP = 0.40, Quality_BRISQUE = 0.35, LPIPS = 0.25); S1 = equal-weight scenario; S2 = semantic-heavy scenario; S3 = aesthetic-heavy scenario; S4 = variation-heavy scenario. Shannon Entropy was excluded from all main AIS-R sensitivity scenarios and retained only as a diagnostic metric.

5. Discussion

The findings indicate that the four generative AI platforms exhibit distinct metric profiles under the APD-50 prompt set and the revised AIS-R configuration. However, these profiles should be interpreted as descriptive comparisons of selected visual-output characteristics rather than as direct evidence of architectural quality, constructability, spatial validity, or professional workflow suitability. The selected computer vision metrics describe image-level properties, including prompt alignment, reference-free perceptual quality, diagnostic information density, intra-prompt perceptual variation, and intra-prompt structural similarity, but they do not directly measure architectural design value. Midjourney v8.1 achieved the highest AIS-R score under the revised configuration and showed relatively higher intra-prompt perceptual variation. This result indicates a stronger performance within the specific composite metric structure used in this study, but it should not be interpreted as evidence that Midjourney produces architecturally superior or more meaningful design alternatives. Rather, it suggests that under the APD-50 prompt conditions, Midjourney generated outputs with comparatively higher combined values across the selected AIS-R components. Stable Image Ultra showed relatively strong prompt alignment and intra-prompt structural similarity within the evaluated metric set. However, these values should not be interpreted as evidence of higher facade correctness, material authenticity, constructability, or spatial logic. Similarly, OpenAI Image Generation exhibited comparatively low Raw BRISQUE values, indicating fewer reference-free perceptual distortions under the selected preprocessing and evaluation conditions, but this does not directly imply superior suitability for professional presentation contexts. Imagen 4 showed strong diagnostic information-density values and competitive AIS-R scores in several categories; however, because Shannon Entropy was excluded from AIS-R, higher information density should not be interpreted as superior architectural detail, material richness, or design quality. Overall, the results should be read as a controlled metric-based characterization of generated image outputs rather than as a professional assessment of architectural design performance. The observed metric profiles may generate hypotheses for future workflow-based studies, such as whether certain platforms are more useful in early-stage visual exploration, facade and material studies, or conceptual presentation

contexts. However, such workflow associations were not directly tested in the present experiment and therefore should not be treated as validated conclusions. Future studies should examine these possible relationships through architect-led evaluations, controlled user studies, and real architectural design workflow experiments.

Limitations

This study has several limitations.

1. Lack of professional architectural jury validation: AIS-R is not a validated measure of architectural quality, constructability, spatial validity, scale accuracy, or professional design value; it is a study-specific metric-based summary of selected visual-output characteristics. No professional jury of architects or design experts evaluated the generated images.
2. Prompt dataset scope: APD-50 consists of 50 prompts across five typologies and should not be interpreted as a universal or exhaustive benchmark of architectural visualization. Model rankings may depend on the specific composition, wording, and category structure of the prompt set.
3. Sample size per model–prompt combination: Only four images were generated per model–prompt combination. Therefore, LPIPS and SSIM values should be interpreted as local indicators of intra-prompt perceptual variation and structural similarity rather than exhaustive measures of generative behavior.
4. Closed commercial platform opacity: OpenAI Image Generation, Imagen 4, Midjourney v8.1, and Stable Image Ultra are commercial systems with limited transparency regarding internal model versions, seed values, post-processing, and inference parameters. Observed differences may therefore reflect model-interface configurations rather than underlying model architecture alone.
5. Aspect ratio and resizing effects: Although preprocessing was standardized prior to metric calculation, raw outputs differed in native resolution and aspect ratio. Resizing may introduce compositional or perceptual effects that cannot be fully eliminated.
6. Heuristic metric weighting: AIS-R weights are study-specific heuristic choices and are not derived from expert elicitation, AHP, Delphi procedures, or empirical calibration against professional architectural judgments. Sensitivity analysis showed that rankings were stable across several scenarios but partially sensitive under variation-heavy weighting.
7. Untested workflow implications: The discussion of possible workflow relevance is theoretical and derived from metric profiles. The study did not directly test model performance in active architectural design workflows, client presentations, professional studios, or architectural education settings.

These limitations directly constrain the interpretation of AIS-R and should not be treated as issues that are fully resolved by acknowledgement alone. Accordingly, AIS-R should not be used as a final or externally validated architectural evaluation framework. Its contribution in the present study is limited to a first-stage, transparent, and reproducible operational formulation for summarizing selected image-level characteristics under controlled prompt conditions. The validity of this formulation for professional architectural judgment remains to be tested in subsequent studies involving architect-led evaluation, expert-calibrated weighting, and workflow-based validation.

6. Conclusions

This study presented a quantitative metric-based comparison of four generative AI platforms for architectural visualization using the APD-50 controlled prompt set and an 800-image corpus. In response to methodological concerns regarding the interpretation of

image-level metrics and heuristic composite weighting, the original AIS formulation was revised as AIS-R. The resulting AIS-R excludes Shannon Entropy and combines normalized CLIP, Quality_BRISQUE, and LPIPS Perceptual Variation as an explicit and reproducible first-stage composite index for summarizing selected visual-output characteristics under controlled prompt conditions. The findings indicate that the evaluated platforms exhibit different metric profiles across prompt alignment, reference-free perceptual quality, intra-prompt perceptual variation, and intra-prompt structural similarity. Under the revised AIS-R configuration, Midjourney v8.1 achieved the highest composite score, followed by Stable Image Ultra, Imagen 4, and OpenAI Image Generation. Linear mixed-effects modeling and prompt-level aggregated robustness analysis indicated statistically significant model fixed effects. However, these findings should be interpreted as metric-based differences within the APD-50 experimental setup rather than as definitive evidence of architectural quality, constructability, spatial validity, scale accuracy, or professional workflow suitability. The main contribution of this study is therefore not the validation of a final architectural-quality score, but the operational formulation and controlled benchmarking of an exploratory composite metric for AI-generated architectural visualization outputs. AIS-R should be understood as a transparent and study-specific starting formulation that enables selected image-level characteristics to be summarized and compared in a reproducible way. Its weighting scheme remains heuristic, and its correspondence with professional architectural judgment has not yet been externally validated. Accordingly, this study represents the first stage in a broader AIS-R development pathway: metric formulation and controlled benchmark testing. Future research should constitute the second stage by evaluating the correspondence between AIS-R and structured assessments by professional architects, architectural educators, or design juries. Such validation studies should examine whether AIS-R components and weights correlate with expert judgments of architectural quality, spatial coherence, material plausibility, scale accuracy, constructability, and workflow usefulness. Future work should also incorporate larger and more diverse prompt sets, alternative weighting strategies, user studies, and real workflow-based experiments to determine the practical relevance and external validity of the proposed metric formulation.

Author Contributions: Conceptualization, M.K.; methodology, M.K., E.C. and V.C.; software, E.C. and V.C.; validation, E.C. and V.C.; formal analysis, E.C. and V.C.; investigation, M.K., E.C. and V.C.; resources, M.K.; data curation, E.C. and V.C.; writing—original draft preparation, M.K., E.C. and V.C.; writing—review and editing, M.K., E.C. and V.C.; visualization, E.C. and V.C.; supervision, M.K.; project administration, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Istanbul Topkapi University within the scope of its publication incentive program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used artificial intelligence tools for language editing, translation, and assistance in code development. The authors have reviewed and edited the outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Complete APD-50 Prompt Set

A. Modern Residential Buildings (P01–P10)	P01	Modern two-story villa, minimalist architecture, wooden facade, large glass windows, sustainable materials.
	P02	Minimalist concrete house, green roof, floor-to-ceiling windows, natural lighting, sustainable architecture.
	P03	Luxury residential villa, modern architectural style, swimming pool, elegant facade, modern landscape design.
	P04	Contemporary smart home, minimalist architecture, solar panels, sustainable materials, energy-efficient design.
	P05	Modern urban townhouse, contemporary architecture, glass balconies, clean geometric lines, minimalist facade.
	P06	Scandinavian style villa, wooden textures, large garden, warm atmosphere, natural materials.
	P07	Futuristic luxury house, innovative architecture, curved forms, dynamic facade design, high-tech materials.
	P08	Three-story residential building, modern architecture, natural stone facade, large windows, sustainable design.
	P09	Eco-friendly family house, green architecture, integrated with nature, sustainable materials, energy-efficient design.
	P10	Contemporary lakeside villa, wooden facade, panoramic windows, modern architecture, natural environment.
B. Public Buildings (P11–P20)	P11	Modern university campus building, contemporary architecture, open courtyard, glass facade, landscaped public spaces.
	P12	Public library, modern architecture, glass facade, spacious reading halls, sustainable design.
	P13	Modern city hall building, contemporary architecture, transparent glass facade, public plaza, civic urban design.
	P14	Futuristic museum building, organic architectural forms, innovative facade design, contemporary architecture, iconic landmark design.
	P15	Modern hospital building, healing architecture, green spaces, large glass windows, patient-centered design.
	P16	Large convention center, parametric roof structure, modern architecture, wide entrance plaza, glass facade.
	P17	Contemporary airport terminal, modern architecture, spacious interior, glass curtain wall facade, dynamic structural design.
	P18	Modern science center, contemporary architecture, interactive public spaces, innovative facade, educational building.
	P19	Educational complex, sustainable architecture, open courtyards, modern academic buildings, green campus design.
	P20	Modern government building, monumental entrance, elegant facade, contemporary civic architecture, public square.

C. Religious and Cultural Architecture (P21–P30)	P21	Modern mosque, contemporary Islamic architecture, large central dome, elegant minarets, ornamental courtyard.
	P22	Traditional mosque courtyard, modern architectural interpretation, Islamic geometric patterns, stone arcades, open prayer space.
	P23	Islamic cultural center, sustainable architecture, open public spaces, modern facade design, green landscape integration.
	P24	Contemporary church, minimalist architecture, natural daylight, large stained glass windows, peaceful interior atmosphere.
	P25	Modern temple, contemporary architecture, natural landscape integration, wood and stone materials, harmonious spatial composition.
	P26	Cultural center, Ottoman-inspired architecture, traditional arches, courtyard design, modern material interpretation.
	P27	Modern exhibition hall, contemporary architecture, Islamic geometric facade patterns, glass and steel structure, large public exhibition space.
	P28	Contemporary memorial building, symbolic architectural forms, minimalist monument design, reflective materials, public gathering space,
	P29	Traditional Turkish cultural house, modern architectural interpretation, wooden facade, courtyard garden, Ottoman-inspired details.
	P30	Modern prayer hall, innovative roof structure, daylight integration, minimalist sacred architecture, calm spiritual atmosphere.
D. Interior Design (P31–P40)	P31	Modern living room, minimalist interior design, wooden furniture, large windows with natural lighting, warm neutral color palette.
	P32	Luxury hotel lobby, contemporary interior architecture, marble flooring, double-height ceiling, elegant reception area.
	P33	Minimalist office interior, open workspace design, biophilic architecture, natural wood materials, large indoor plants.
	P34	Modern kitchen interior, minimalist design, smart appliances, sustainable materials, functional storage solutions.
	P35	Contemporary bedroom, modern interior design, natural wood textures, soft ambient lighting, relaxing atmosphere.
	P36	Luxury restaurant interior, contemporary design, elegant lighting fixtures, premium materials, stylish dining atmosphere.
	P37	Modern university classroom, flexible learning environment, smart educational technologies, modular furniture, large collaborative learning spaces.
	P38	Art gallery interior, minimalist exhibition design, open gallery spaces, white walls, professional exhibition lighting.
	P39	Modern co-working space, open-plan office design, collaborative work areas, glass partition walls, ergonomic furniture.
	P40	Contemporary cafe interior, warm atmosphere, sustainable wooden furniture, large panoramic windows, modern decorative elements.

E. Facade and Sustainable Architecture (P41–P50)	P41	Parametric building facade, dynamic geometric patterns, sustainable materials, contemporary architecture, computational design.
	P42	Biophilic building facade, vertical gardens, natural vegetation integration, sustainable architecture, living green walls.
	P43	Smart building facade, adaptive shading system, energy-efficient architecture, responsive facade technology, modern architectural design.
	P44	Green building facade, photovoltaic solar panels, sustainable materials, contemporary architecture, energy-efficient building envelope.
	P45	Contemporary building facade, perforated metal panels, glass curtain wall, minimalist architecture, modern urban design.
	P46	Responsive building envelope, adaptive facade system, environmental responsiveness, smart materials, innovative architectural design.
	P47	Sustainable office building facade, double-skin glass system, high-performance facade, energy-efficient design, modern office architecture.
	P48	Innovative architectural facade, organic architectural forms, computational design, parametric geometry, advanced building materials.
	P49	Modern mixed-use building, dynamic facade design, green terraces, contemporary urban architecture, glass and sustainable materials.
	P50	Future-oriented skyscraper facade, intelligent shading systems, sustainable high-rise architecture, advanced facade technology, energy-efficient building design.

References

1. Tang, X.; Windham, J.; Bush, B. Pre-AI and post-AI design: Balancing human creativity and AI tools in the industrial design process. In *Proceeding of the 2024 International Conference on Artificial Intelligence and Future Education, Guangdong, China, 1–2 November 2024*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 100–108.
2. Jin, S.; Tu, H.; Li, J.; Fang, Y.; Qu, Z.; Xu, F.; Liu, K.; Lin, Y. Enhancing architectural education through artificial intelligence: A case study of an AI-assisted architectural programming and design course. *Buildings* **2024**, *14*, 1613. [CrossRef]
3. Buldaç, M. Use of artificial intelligence tools in the experimental design process: Outcomes of a course model in interior design education. *J. Art. Des.* **2024**, *14*, 69–91.
4. Avinç, G.M. The use of text-to-image generation artificial intelligence tools for the production of biophilic design in architecture. *Black Sea J. Eng. Sci.* **2024**, *7*, 641–648.
5. Pena, M.L.C.; Carballal, A.; Rodríguez-Fernández, N.; Santos, I.; Romero, J. Artificial intelligence applied to conceptual design: A review of its use in architecture. *Autom. Constr.* **2021**, *124*, 103550. [CrossRef]
6. Bölek, B.; Tutal, O.; Özbaşaran, H. A systematic review on artificial intelligence applications in architecture. *J. Des. Resil. Archit. Plan.* **2023**, *4*, 91–104. [CrossRef]
7. Harapan, A.; Indriani, D.; Rizkiya, N.F.; Azbi, R.M. Artificial intelligence in architectural design. *Int. J. Des. INJUDES* **2021**, *1*, 1–6. [CrossRef]
8. Wen, M.; Liang, D.; Ye, H.; Tu, H. Architectural facade design with style and structural features using stable diffusion model. *J. Intell. Constr.* **2024**, *2*, 1–12. [CrossRef]
9. Zhu, J.Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. *arXiv* **2017**, arXiv:1711.11586. [CrossRef]
10. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef] [PubMed]
11. Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7514–7528.
12. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]

13. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 586–595.
14. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
15. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 4th ed.; Pearson: London, UK, 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.