architecture

MDPI

Article

The Prompt Richness Index: A Proposed Seven-Dimension Framework for Evaluating AI Text-to-Image Generation in Architectural Design Education

Nahedh Taha Al-Qemaqchi [1,2]

[1] Department of Architectural Engineering, Cihan University—Sulaimaniya, Sulaimaniya 46001, Kurdistan Region, Iraq; nahith.taha@sulicihan.edu.krd
[2] Department of Architectural Engineering, Tishk International University—Sulaimaniya, Sulaimaniya 46001, Kurdistan Region, Iraq

Abstract

Transforming concepts into architectural designs is challenging, as it requires clear ideas and the ability to visualise them. AI-powered text-to-image tools help designers quickly convert concepts into visual representations, enabling faster exploration of design alternatives. This study introduces a proposed ten-step numerical procedure for assessing the richness of textual prompts submitted to text-to-image generative AI tools within an architectural design studio. Twenty-three architecture students were enrolled in the design studio; twenty-two submitted analyzable text prompts as part of a design assignment requiring AI-assisted conceptual visualisation. Each prompt was scored across seven weighted dimensions (subject specificity, style and medium, composition and framing, lighting and atmosphere, colour and palette, quality modifiers, and negative clauses) to produce a composite Prompt Richness Index (R, scale 0–100). Corresponding AI-generated images were independently scored using a parallel Output Richness Index (O, scale 0–100). Pearson's r between per-student average R and O yielded r = 0.940 ($p < 0.001$, 95% confidence interval [0.86, 0.98]), confirming a nearly perfect positive linear relationship. Rich-tier prompts were produced by two students and yielded the most architecturally coherent and visually distinctive outputs. Two students produced Sparse-tier prompts (average R < 30) and consistently received undifferentiated, generically rendered outputs. Two further students scored just above the Sparse threshold but showed similarly limited output differentiation. Class-wide deficits were identified in lighting/atmosphere description and negative clause usage. Eight pedagogical recommendations are derived from the findings to guide prompt learning instruction in AI-integrated design studios.

Keywords: prompt engineering; text-to-image AI generation; Prompt Richness Index; Output Richness Index; architectural education

Check for updates

Academic Editor: Avi Friedman

Received: 29 June 2026
Revised: 27 August 2026
Accepted: 31 August 2026
Published: 2 September 2026

Copyright: © 2026 by the author. 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.

1. Introduction

The conceptual design phase is often the most demanding, requiring multiple alternatives and the selection of an optimal solution to express vague ideas. Central to this phase is concept generation, which refers to the creation of ideas to address a design problem [1]. It sets the overall direction, linking abstract thinking to feasible solutions [2]. Ideation here demands cognitive flexibility and problem-solving, while creativity is best understood as a structured, learnable process integrating existing knowledge with new insights rather than sudden inspiration [3,4].

Despite the importance of concept development, conventional architectural education (typically centred on the design studio model) faces significant challenges in fostering effective conceptual thinking [5]. Preliminary architectural design courses often emphasise fundamental design skills and drawing techniques, while giving inadequate attention to the application of basic design concepts or the framing of students' individual creativity [6]. Within the studio framework, students tend to rely heavily on the expertise and experience of their instructors to acquire professional skills [7]. This approach limits opportunities for students to construct or develop systematic knowledge independently [8]. As a result, learners may adopt a passive role, receiving information rather than actively engaging with it, which can hinder inspiration and the development of independent thinking or a clear understanding of the objectives underlying design tasks [9].

Throughout history, the tools and media used for architectural design and representation have had a profound influence on architectural thought and practice. Initially, digital tools such as computer-aided design (CAD) were adopted primarily to increase drafting efficiency [5]. Over time, however, these technologies have reshaped design methodologies and expanded the scope of architectural exploration [10,11]. The introduction of advanced computational techniques, including parametric design, has enabled architects to move beyond simple geometric forms and engage with complex, dynamic design processes [12].

More recently, artificial intelligence (AI) has emerged as a transformative force within the architectural field. AI applications (particularly text-to-image diffusion models) are increasingly integrated into architectural design workflows [13]. These technologies contribute to various aspects of architectural practice, including conceptual design, visual representation, and data-driven analysis [14]. At the same time, the growing presence of AI in architectural education is shifting pedagogical priorities away from purely technical or traditional skills, toward greater emphasis on conceptual thinking, critical reflection, and collaborative design processes [15]. The emergence of text-to-image artificial intelligence (AI) tools has introduced a fundamentally new mode of visual ideation into architectural design practice and education [16]. These systems translate natural language descriptions into architectural imagery, enabling designers to explore their output alternatives at a speed previously unattainable. In studio education, this capability is already being incorporated into early-stage conceptual design and formal exploration, with architecture students using AI-generated imagery as both inspiration and design output [17].

However, the visual quality and conceptual reliability of AI-generated images are not inherent to the generator alone; they depend heavily on the content, structure, and linguistic precision of the input prompt. Different inputs produce varied outputs, and as prompts become more refined, the resulting images become richer and more coherent [18]. Despite the widespread accessibility of generative tools, which has attracted significant interest in architectural education, this critical aspect remains neglected in academic practices. Consequently, design pedagogy still lacks a systematic and replicable framework to measure and teach this variation effectively.

This paper addresses that gap by answering the following three research questions:

1. To what extent does prompt richness affect the quality of AI-generated visual outputs in the architectural concept design phase?
2. What factors determine the richness of prompts in AI-driven image generation, and what is the relative impact of each factor on the quality of the generated output?
3. Finally, what role can prompt richness and prompt engineering play in shaping pedagogical approaches in architectural design education?

To address the above questions, the research adopts a three-fold methodology. First, it proposes a ten-step numerical framework for evaluating prompt richness across seven

weighted dimensions, resulting in a composite Prompt Richness Index (R). Second, the framework is applied to a dataset of prompts produced by architecture students, generating both dimension-specific and overall richness scores. Third, these scores are correlated with independently assessed Output Richness Indices (O), derived from the corresponding AI-generated images. It is important to clearly define the scope of the framework proposed in this study. The Prompt Richness Index (PRI) is not intended to replace the open-ended context-dependent process of architectural concept generation. Instead, it addresses a more manageable task within that broader process: translating an emerging design intention into a textual description suitable for text-to-image generation. In this context, the PRI serves as a diagnostic and pedagogical tool that supports the development of prompt-writing skills, rather than a comprehensive measure of design quality or conceptual depth.

2. Background

2.1. Concept Generation in Architecture Studio

Concept generation is vital in the architectural design process, as it provides the underlying thematic inference required for a successful proposal [19]. In the contemporary design studio, design has moved beyond a sole focus on appropriateness and utility, placing equal emphasis on developing and realising new ideas [20]. This fundamental stage focuses on the methodologies of conceptualisation, including the derivation, processing, and translation of preliminary concepts into a concrete design direction. A holistic approach to concept evolution seeks to unify the many elements of design conception into a framework that enables integration throughout different phases of the project [21].

Many scholars assert that the initial stages of architectural design present difficulties for inexperienced students, especially in developing abstract conceptions and correlating them with spatial arrangements [22–24]. Conventional methods frequently depend significantly on sketching, prompting students to uncover ideas spontaneously. Nevertheless, the reliance on sketching disadvantages novices who do not have the organised conceptual frameworks that specialists do [25]. These sketches serve as externalised cognitive frameworks that can be progressively enhanced, promoting intellectual growth within the studio setting [26]. Language, as a parallel medium, allows students to verbalise abstract concepts, making their design intentions explicit and communicable [27,28]. The combined use of verbal and visual media enhances reflective practice and strengthens the designer's cognitive processes [27].

Taura and Nagai classify concept generation into two complementary phases [1]:

- The Problem-Driven Phase entails reacting to clear requirements such as site limits, programming needs, or environmental considerations. The architect examines the situation to find solutions, such as using solar path analysis to determine building orientation.
- Inner Sense-Driven Phase: This is guided by the designer's internal standards, ideals, and motives. It is where abstract concepts like "transparency," "layering," or "resilience" arise, frequently inspired by personal or societal narratives, metaphors, or creative influences.

In architecture, these stages are rarely separated. A notion may come from an internal feeling of "flow" or "connection," which is subsequently evaluated and improved upon considering real-world limitations. It is common for innovative architectural ideas to emerge from a wider "thought space" [29]. This can be promoted in the studio by employing strategies like mind mapping, analogy, or interdisciplinary references, which go beyond traditional typologies to uncover fresh spatial and experiential opportunities [30].

2.2. Artificial Intelligence in Architecture

The conceptual roots of AI in architecture can be traced back to the mid-twentieth century, when pioneers began to investigate computation as a design tool. According to [31], the experimental legacy of computational aesthetics and design serves as the foundation for present AI approaches to architecture. Early efforts, such as Nicholas Negroponte's Architecture Machine and Soft Architecture Machines, postulated "thinking machines" capable of learning and responding to human input ideas, paving the way for participatory and adaptive design systems [31].

These initial efforts were shaped by cybernetics, a forerunner of contemporary machine learning, which examined information feedback and adaptation. In Design for a Brain, researchers such as Ross Ashby examined the simulation of neurons and adaptive behaviours, predicting the emergence of artificial neural networks long before they were computationally feasible. In the 1980s, shape grammar arose as an algorithmic system for creating architectural forms [32] from a set of transformation rules, representing one of the first instances of rule-based AI in design [33].

By the 1990s, architects like John and Jane Frazer had expanded these concepts into "evolutionary architecture," which used genetic algorithms that simulated biological evolution to investigate design possibilities. Such advancements marked a transition from deterministic computation to adaptive and generative models that more closely resembled human creativity [32].

In recent decades, the maturation of AI and ML has revolutionised architectural workflows. The computer is no longer merely "a faster pencil," but an active participant in creative and analytical processes. The evolution from computer-aided design (CAD) to building information modelling (BIM) and parametric design tools paved the way for data-driven and AI-integrated practices [17].

2.3. Text-to-Image Generative AI

Diffusion-based text-to-image models represent the current state of the art in generative image synthesis. Rombach et al. (2021) introduced latent diffusion models that operate in a latent space, enabling high-resolution synthesis at reasonable computational cost [34]. Saharia et al. (2022) demonstrated that deep language understanding (encoding not just keywords but semantic relationships and contextual meaning) substantially improves text-image alignment in generation [35]. These architectural advances mean that modern generators are sensitive to precise prompt phrasing, not merely keyword presence.

Tools such as DALL-E, Midjourney, Stable Diffusion, and Leonardo AI demonstrate how text-to-image diffusion models can assist architects and designers in rapidly visualising concepts during early design stages [36]. These image generators accept natural-language prompts and return visual outputs. Complementary to these, large language models (LLMs) such as ChatGPT, Microsoft Copilot, and Gemini do not generate images but can support the prompt-writing process itself by helping users articulate, refine, and expand their textual descriptions before submitting them to an image generator [13].

Despite this sensitivity, the relationship between prompt specificity and output quality has received limited quantitative empirical treatment. Most published work on prompt quality is descriptive, taxonomic, or user-study-based rather than numerically procedural and correlational.

A related body of scholarship focuses less on the technical effectiveness of prompts and more on the institutional and pedagogical expectations surrounding their documentation and disclosure [10]. Recent studies have advocated higher education policies that require students to acknowledge their use of generative AI tools and, within design disciplines,

to preserve and disclose the prompts underlying submitted visual outputs [11,13]. The prompt-richness framework proposed in this study, together with the end-of-semester prompt portfolio outlined in Section 6, aligns with and supports this emerging norm of transparency. By requiring students to systematically record, curate, and critically reflect on the prompts used throughout a project, the framework provides a practical mechanism for operationalising responsible AI disclosure practices.

2.4. Prompt Engineering: Taxonomies and Modifiers

Oppenlaender (2022) proposed a taxonomy of prompt modifiers for text-to-image systems, identifying five principal categories: subject terms, style tokens, quality boosters, composition descriptors, and negative prompts [37]. Han & Fussell (2025) demonstrated through user perception studies that prompt length and specificity correlate with user-rated satisfaction in AI image generation tasks but did not develop a formal scoring instrument or test the correlation against independently evaluated output quality [38]. Rombach et al. (2021) identified descriptor categories in text-to-image prompts and showed that style modifiers had the strongest effect on output diversity [34].

The proposed framework builds upon Oppenlaender's taxonomy by introducing theoretically grounded weights for each dimension and integrating the seven dimensions into a unified composite index. This approach enables a more systematic assessment of prompt richness and provides a quantitative measure suitable for statistical analysis and empirical research. These insights support the differential weighting scheme applied in the Prompt Richness Index. Subject specificity and style/medium receive the highest weights ($w_1 = w_2 = 2$) because these dimensions exert the strongest influence on the generator's compositional output and visual character [34,37]. Composition and lighting receive intermediate weights ($w_3 = w_4 = 1.5$) based on their documented role in spatial framing [34,37]. The remaining dimensions (colour, quality modifiers, negative clauses) receive equal lower weights ($w_5 = w_6 = w_7 = 1$), reflecting their supporting rather than primary role in image generation. These weights represent theoretically motivated expert assignments; empirical calibration through a larger multi-platform dataset is identified as a direction for future work. We emphasise that this weighting scheme reflects theoretically motivated expert judgment rather than empirically calibrated or psychometrically validated weights; no claim of formal instrument validation (e.g., construct, convergent, or criterion validity) is made at this stage.

3. Materials and Methods

3.1. Study Context and Participants

The study was conducted within an architectural design studio at Tishk International University in Sulaimaniya. Twenty-three third-stage students (labelled S1–S23, with S8 absent from the dataset because the student submitted images without an accompanying text prompt) were assigned the task of developing conceptual visualisations of a riverside restaurant located on a sloped site with an elevation change of 2–8 m. The site context (a sloping river edge) served as a fixed constraint, providing a common design brief against which prompt specificity could be evaluated. The experimental design session lasted for one week, allowing students sufficient time to iteratively refine their generated images before submission. Students participated within the normal scope of the studio assignment and were informed that their submitted prompts and generated images would be used for educational research purposes.

Students were free to use any text-to-image AI generator of their choice. No prior instruction on prompt writing was provided, as the study was intentionally designed to capture participants' baseline prompt literacy before any pedagogical intervention. While

different text-to-image models, such as Midjourney V7, DALL·E 3, Stable Diffusion 3.5, and Leonardo AI V2, may produce varying outputs in response to the same prompt, this variable was deliberately held outside the scope of the investigation to maintain focus on prompt formulation rather than model-specific performance. Each student submitted between one and four prompts, along with the corresponding AI-generated images. In total, 55 prompts and 55 associated images were collected and analysed.

3.2. The Ten-Step Prompt Richness Procedure

The core methodological approach of this study is a ten-step procedure for numerically evaluating prompt richness. The procedure is reproducible, dimension-specific, and produces a single composite index (R) on a 0–100 scale. Table 1 presents the full step-by-step procedure.

Table 1. The Ten-Step Prompt Richness Evaluation Procedure.

Step	Dimension/Action	Procedure
1	Subject specificity (S_1)	Count and classify main subjects. Score on a 0–3 scale based on precision: 0 = absent; 1 = generic noun; 2 = named or described; 3 = named with a full set of attributes.
2	Style and medium (S_2)	Identify style-encoding tokens. Score on a 0–3 scale: 0 = absent; 1 = broad genre; 2 = named artist or movement plus medium; 3 = style, medium, and historical era combined.
3	Composition and framing (S_3)	Count spatial framing cues (e.g., shot type, camera angle, depth references). Assign one point per distinct cue, up to a maximum score of 3.
4	Lighting and atmosphere (S_4)	Identify lighting characteristics, including source type, quality, and mood. Score on a 0–3 scale: 0 = absent; 1 = basic mood; 2 = clearly specified light source; 3 = fully articulated, layered atmospheric lighting.
5	Colour and palette (S_5)	Check for explicit references to colours, palette terminology, and cues related to saturation and contrast. Score on a 0–3 scale: 0 = absent; 1 = general mention; 2 = defined or named palette; 3 = fully articulated colour scheme or colour story.
6	Quality and rendering modifiers (S_6)	Count technical quality tokens (e.g., 8K, photorealistic, Octane render). Score on a 0–3 scale: 0 = none; 1 = one modifier; 2 = two modifiers; 3 = three or more modifiers.
7	Negative/exclusion clauses (S_7)	Check for exclusion terms. Score on a 0–2 scale: 0 = absent; 1 = vague exclusion; 2 = specific or multi-element exclusions.
8	Compute composite R	Apply weighted formula: $R = (\sum w_i \times S_i)/(3 \times \sum w_i) \times 100$. Default weights: $w_1 = w_2 = 2$; $w_3 = w_4 = 1.5$; $w_5 = w_6 = w_7 = 1$. $R \in [0, 100]$.
9	Classify the richness tier	Sparse: $R < 30$. Moderate: 30–59. Rich: 60–79. Highly Specific: $\geq$80.
10	Diagnose and iterate	Sort dimensions by score ascending. Lowest dimension = highest-leverage revision target. Re-score after revision until the target tier is reached.

Steps 1–7 each correspond to one scoring dimension. Steps 8–10 compute the composite score, classify the prompt into a richness tier, and initiate a diagnostic revision loop. The procedure is designed to be applied iteratively: a student who scores a prompt before generation, identifies the lowest-scoring dimension, adds targeted descriptors, and re-scores will systematically improve output quality through each cycle (Figure 1).

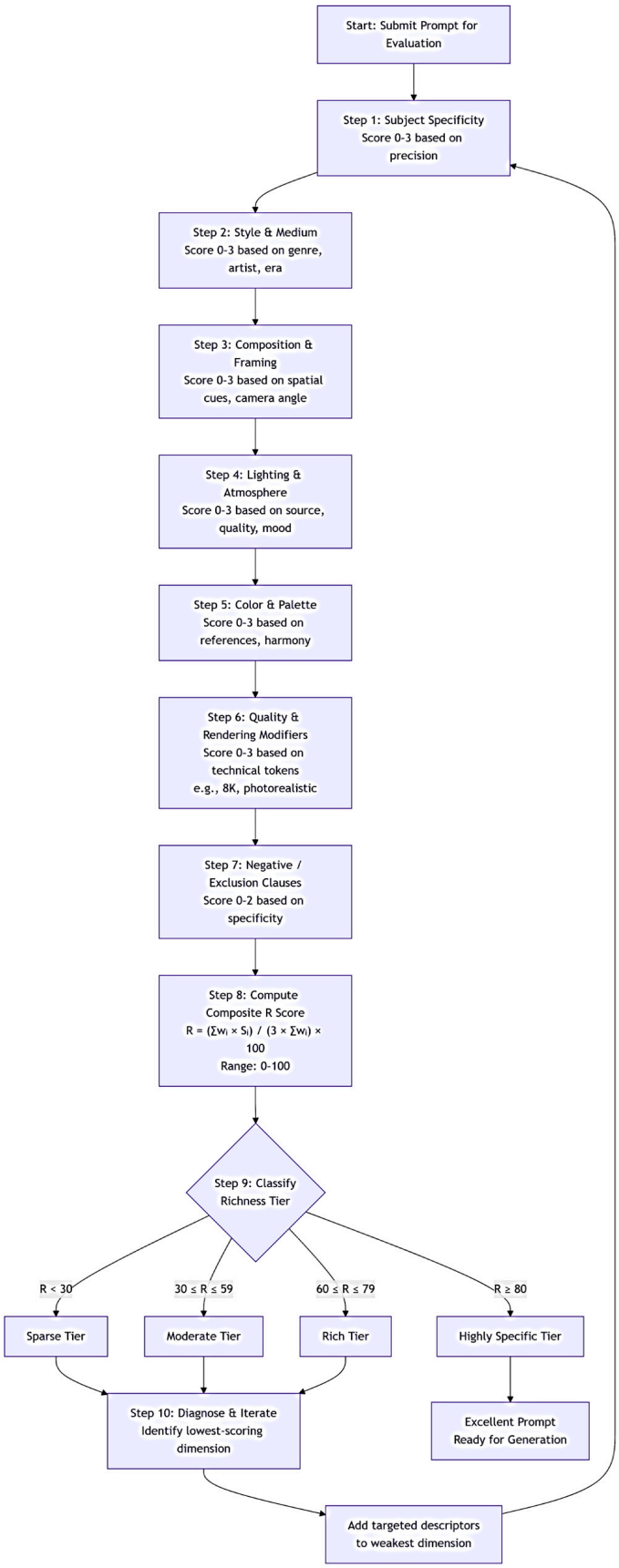


Figure 1. Flow diagram of the Ten-Step Prompt Richness Evaluation Procedure.

3.3. The Scoring Rubric (Dimensions S_1–S_7)

Each dimension is scored on a 0–3 scale (0–2 for S_7) and assigned a weight reflecting its contribution to image generation quality (Table 2).

Table 2. Prompt Richness Scoring Rubric.

Dimension	Wt	Max	Score 0	Score 1	Score 2	Score 3
Subject specificity	2	3	None	Generic noun	Named/described	Named plus attributes
Style & medium	2	3	None	Broad genre	Named style/artist	Style plus medium plus era
Composition & framing	1.5	3	None	One spatial cue	Two cues	Three plus cues
Lighting & atmosphere	1.5	3	None	Basic mood	Named source	Layered lighting
Color & palette	1	3	None	General color	Named palette	Full palette
Quality modifiers	1	3	None	One modifier	Two modifiers	Three + modifiers
Negative clauses	1	2	Absent	Present, vague	Targeted exclusions	N/A

3.4. The Composite Prompt Richness Index (R)

The composite index is computed as a weighted mean of dimension scores, normalised to a 0–100 scale:

$$R = (\sum w_i \times S_i)/(3 \times \sum w_i) \times 100$$

where S_i is the score on dimension i, w_i is its weight, and the denominator ($3 \times \sum w_i$) normalises to the theoretical maximum. With the weights used in this study ($\sum w_i = 10$), the denominator is 30, and R is always in [0, 100] regardless of weight distribution. Four richness tiers were defined: Sparse (R < 30), Moderate (30–59), Rich (60–79), and Highly Specific ($\geq$80). All dimension-score rows (55 prompts $\times$ 2 raters) were re-verified programmatically against the reported R values prior to submission.

Reliability of Prompt Richness Scoring

To assess the reliability of the R scoring procedure, a second rater (trained on the ten-step rubric) independently re-scored all 55 prompts in the dataset. Inter-rater reliability was assessed using the two-way mixed-effects, absolute-agreement intraclass correlation coefficient, ICC(2, k). The same reliability model was applied to the Output Richness Index in Section 3.5.3. ICC values were computed for each of the seven R dimensions individually and for the composite R score (Table 3). All results reported in Sections 4, 5, 6 and Figure 2, and the Pearson correlation r = 0.940) use Rater 1's (the author's) scores as the primary analytic dataset; Rater 2's independent scoring is used exclusively to establish the inter-rater reliability reported in Table 3. Because the author also served as Rater 1, we note this as a potential source of rater bias in Section 6.

Table 3. Inter-Rater Reliability (ICC) for Prompt Richness and Output Richness Dimensions.

Dimension	Weight	ICC (2, k)—R	ICC (2, k)—O
S1/D1 Subject specificity	2	0.90	0.92
S2/D2 Style & medium	2	0.93	[not individually reported]
S3/D3 Composition & framing	1.5	0.88	[not individually reported]
S4/D4 Lighting & atmosphere	1.5	0.66	[not individually reported]

Table 3. *Cont.*

Dimension	Weight	ICC (2, k)—R	ICC (2, k)—O
S5/D5 Colour & palette	1	0.90	[not individually reported]
S6/D6 Quality modifiers	1	0.69	[not individually reported]
S7/D7 Negative clauses/ Conceptual fidelity	1	undefined (zero variance- no negative clauses present in any of the 55 prompts, for either rater)	0.79
Composite R (prompt level, n = 55)/Composite O	-	0.95	0.87 (95% CI 0.76–0.93)
Composite R (student level, n = 22)	-	0.97	[not computed at student level in Section 3.5.3]

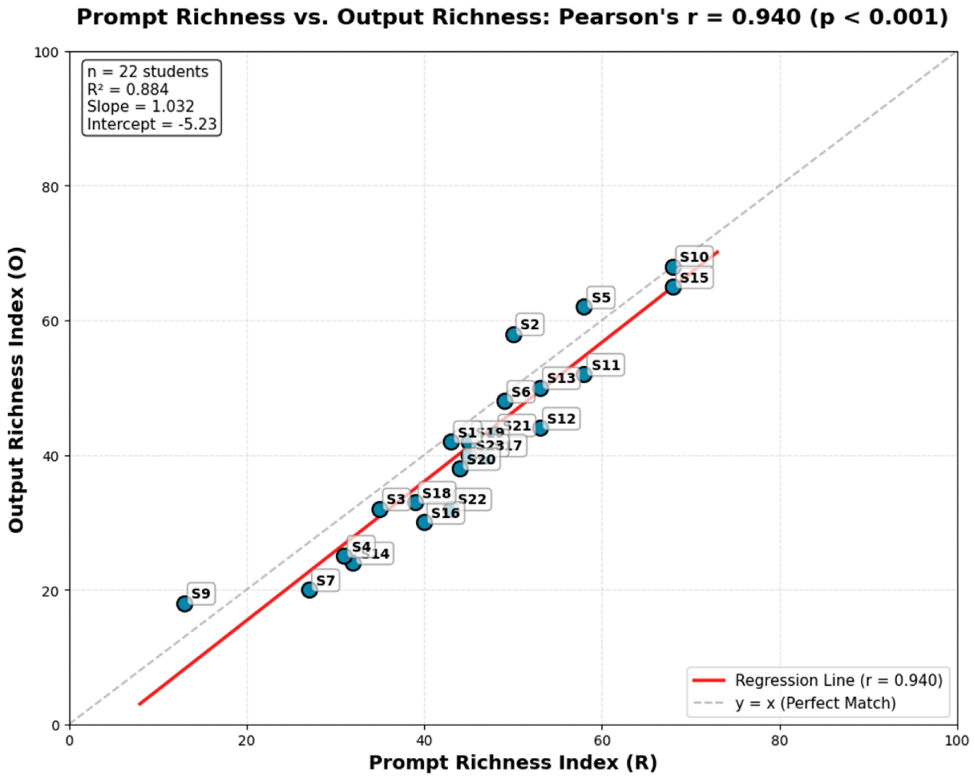


Figure 2. Relationship Between Per-Student Average Prompt Richness (R) and Output Richness (O) (r = 0.940, 95% CI [0.86, 0.98], n = 22).

Two dimensions show only moderate inter-rater reliability by the Koo and Li [39] thresholds cited above: lighting and atmosphere (S4, ICC = 0.66) and quality modifiers (S6, ICC = 0.69), both below the 0.75 threshold for 'good' reliability. Because S4 carries an elevated weight (w = 1.5) and the lighting/atmosphere gap is reported as a headline class-wide deficit (Section 4.3), this moderate reliability should be read as a caveat on the precision, though not necessarily the direction, of that finding.

3.5. Output Richness Scoring (O)

3.5.1. The Output Richness Index (O): Dimensional Structure and Weighting

Each set of AI-generated images was scored using a parallel Output Richness Index (O) across seven analogous visual dimensions (D):

1. Visual subject clarity (D_1); the degree to which the architectural form and programme are clearly identifiable in the output.

2. Style coherence (D_2); the consistency of the visual register (architectural style, artistic medium, and historical references) relative to the prompt intent.
3. Spatial composition quality (D_3); the clarity of depth, framing, camera angle, and spatial hierarchy in the rendered scene.
4. Lighting quality and atmosphere (D_4); the specificity and appropriateness of the lighting condition, including light source, directionality, shadow quality, and overall mood.
5. Colour palette richness and control (D_5); the degree to which a coherent, intentional colour scheme is evident, including named colours, tonal relationships, and saturation control.
6. Rendering detail and technical quality (D_6); the level of material resolution, surface articulation, and technical finish.
7. Conceptual fidelity (D_7); the degree to which the output demonstrably reflects the specific architectural concept, site condition, or programmatic intent communicated in the prompt.

Each dimension was scored on a 0–3 scale (0–2 for dimension 7, conceptual fidelity, to maintain parity with the negative-clause ceiling in the Prompt Richness Index) and combined into O using the same weighted formula as R:

$$O = (\sum w_i \times D_i)/(3 \times \sum w_i) \times 100$$

Weights were assigned to O dimensions to mirror exactly those of the corresponding R dimensions (Table 4). This mirroring ensures that the two indices are directly comparable and that any Pearson's r between R and O reflects genuine alignment between prompt content and visual output, rather than an artefact of asymmetric dimensional weighting. This design choice, however, also means that R and O are not fully independent measurements: because O's dimensions and weights were constructed to track R's structure, part of any observed correlation between the two indices is attributable to this shared design rather than to an independent empirical relationship. This limitation is discussed further in conclusions.

Table 4. Output Richness Scoring Rubric.

Dimension	Wt	Max	Score 0	Score 1	Score 2	Score 3
D_1	2	3	Architectural form absent or unidentifiable; output is an abstract or unrelated scene	Main form partially visible; programme ambiguous; building type guessable but unclear.	Restaurant form clearly identifiable; key elements (entrance, dining zones) legible	Fully resolved architectural subject; programme, scale, and spatial hierarchy all explicit
D_2	2	3	No identifiable visual style; generic or contradictory rendering; no correspondence with prompt	Broad style present (e.g., modern, organic) but inconsistently applied across the image	Named or implied style clearly and consistently rendered; medium and register appropriate	Fully coherent style with named movement, medium, and historical or cultural references faithfully rendered
D_3	1.5	3	No spatial depth or framing; flat, centred, or structureless composition	Basic depth cue or single framing device present; composition readable but generic	Clear spatial hierarchy; foreground, mid-ground, background differentiated; viewpoint intentional	Precise camera angle, shot type, depth layering, and spatial sequence all evident and purposeful

Table 4. *Cont.*

Dimension	Wt	Max	Score 0	Score 1	Score 2	Score 3
D_4	1.5	3	Flat, neutral, or studio-default lighting; no mood or atmospheric quality	Basic ambient mood established; directionality implied but light source indeterminate	Named or identifiable lighting condition (e.g., golden hour, overcast diffuse); consistent shadow quality	Fully layered atmospheric scene: light source, quality, directionality, and mood all specifically and coherently rendered
D_5	1	3	Arbitrary or absent colour palette; colours appear as generator defaults with no intentionality	General colour tone present (e.g., warm, cool) but not controlled or specific	Defined palette with recognisable colour relationships; saturation and contrast intentional	Fully articulated colour story: named or precisely identifiable hues, tonal harmony, and palette control evident throughout
D_6	1	3	Low-resolution or artefact-laden output; materials indeterminate; structural logic absent	Basic material types legible; surface quality present but generic; some structural plausibility	Material properties faithfully rendered; surface articulation clear; structural system plausible.	High-fidelity rendering: material grain, reflectivity, jointing, and openings all resolved; output approaches photorealistic architectural quality.
D_7	1	2	No correspondence between output and prompt concept; generator has completely overridden intent	Core concept partially present; key conceptual element recognisable but architectural resolution weak.	Prompt concept clearly and faithfully expressed in output; site, programme, and formal intent all reflected.	N/A

3.5.2. Judging Panel: Composition, Enrollment, and Independence Protocol

Output scoring was conducted by a panel of three independent judges, all practising design tutors with a minimum of ten years of professional experience in architectural design and design education. Panel members were recruited from outside the research institution to eliminate the risk of familiarity bias with individual students' work. None of the judges had access to the students' prompt texts, R scores, or any other evaluation data before or during the scoring session. This blind evaluation protocol was adopted to ensure that output scores reflected the essential visual properties of the images rather than expectations derived from prompt content.

Judges received a briefing document specifying the design brief (a riverside restaurant on a sloped site of 2–8 m elevation change) and the anchored scoring rubric for all seven O dimensions. Each judge independently evaluated all 23 student image sets and recorded scores for each of the seven dimensions on a standardised scoring sheet. Scoring was conducted asynchronously over five days; judges were instructed not to communicate with one another or with the research team about individual cases until all score sheets had been submitted.

3.5.3. Score Aggregation, Inter-Rater Reliability, and Final O Computation

Final O scores for each student were computed as the arithmetic mean of the three judges' independent composite O scores. This aggregation method was selected for two reasons: the arithmetic mean is the most interpretable central-tendency measure when all raters apply the same ordinal scale to the same object, and it preserves direct comparability with R, which is itself a weighted arithmetic mean of dimension scores. All three judges

contributed equally to the final O value; no judge's score was excluded, downweighted, or adjusted before aggregation.

Inter-rater reliability (IRR) was assessed post hoc using the intraclass correlation coefficient (ICC), specifically the two-way mixed-effects model for absolute agreement, ICC(2, k). This model is appropriate when judges represent a fixed panel (rather than a random sample of all possible raters) and when the research objective is to establish agreement on absolute score values [39]. ICC values were computed for each of the seven O dimensions individually and for the composite O score. The composite ICC was 0.87 (95% CI: 0.76–0.93, ICC computed at the student/portfolio level), indicating good-to-excellent reliability according to the thresholds of Koo and Li [39]. Dimension-level ICC values ranged from 0.79 (conceptual fidelity, D_7; the most subjective dimension) to 0.92 (visual subject clarity, D_1; the most objectively verifiable). This gradient is consistent with theoretical expectations: dimensions that can be assessed against explicit visual evidence (clarity of form, material rendering) attract higher inter-judge agreement than dimensions requiring interpretive judgement (conceptual fidelity, atmospheric coherence). The composite ICC of 0.87 is sufficient to justify the use of the mean O score as a stable and trustworthy measure for Pearson's r analysis.

In the calibration session, two exemplary image sets drawn from a separate student cohort (not included in the study) were scored independently and then discussed until consensus was reached on the interpretation of each anchor. No modifications to the rubric were made following calibration; the session served exclusively to align judges' understanding of the anchors without introducing familiarity with the study's main dataset.

Judges were additionally instructed to score each of the seven dimensions independently before computing any composite impression, to minimise halo effects (the tendency for an initial holistic reaction to an image to inflate or deflate scores on individual dimensions [40]). Where a student submitted multiple images, judges evaluated the full set holistically, assigning a single set of seven-dimension scores reflecting the overall visual richness of the student's AI output portfolio. This portfolio-level aggregation is consistent with the per-student unit of analysis adopted throughout the study.

4. Results

4.1. Full Dimension Scores and R Values

Per-student average R, computed as the arithmetic mean of R scores across a student's 1–4 submitted prompts, was correlated with a single student-level portfolio O score using Pearson's r. The Pearson's r coefficient was chosen as the primary statistical measure because both R and O have continuous numeric indices on the same 0–100 scale, the relationship was hypothesised to be linear, and the sample (n = 22) is sufficient for a Pearson analysis at $\alpha = 0.05$. The significance threshold was set at $p < 0.05$. It should be noted that per-student average R (computed across all their prompts) is correlated with a single student-level O score (reflecting the overall visual richness of their image portfolio). This aggregation approach was adopted because students submitted varying numbers of prompts (1–4) and because output quality was assessed holistically per student rather than per image. A prompt-image paired analysis was beyond the scope of this study but is identified as a methodological refinement for future work (Table 5).

Table 5. All 55 Prompts with Dimension Scores (S1–S7) and Composite Richness Index (R) for the two raters.

No.	Student	Prompt Text	Rater-1 Scores		Rater-2 Scores	
			S1, S2, S3, S4, S5, S6, S7	R	S1, S2, S3, S4, S5, S6, S7	R
1	S1–P1	A fluid, organic restaurant inspired by the gentle flow of water; its curving forms mimic waves shaping the shore, blending architecture and nature.	2, 2, 1, 1, 1, 1, 0	43	2, 1, 0, 1, 0, 3, 0	35
2	S1–P2	A riverside restaurant with a parasitic-inspired design, featuring a ribbed wooden extension that clings to the main stone structure.	2, 2, 1, 1, 1, 1, 0	43	2, 2, 1, 0, 1, 1, 0	38
3	S1–P3	A bold parasitic form anchored in stone, the design emerges from the landscape, blending raw nature with sharp geometry in harmony.	2, 2, 1, 1, 1, 1, 0	43	2, 3, 0, 0, 0, 1, 0	37
4	S2–P1	A futuristic sustainable restaurant located on an 8-m slope beside a river, designed with fluid, wave-like layers inspired by ocean movement and intersected with sharp crystal geometry.	3, 3, 2, 1, 1, 1, 0	62	3, 3, 2, 0, 1, 1, 0	57
5	S2–P2	A futuristic restaurant inspired by ancient Mesopotamia, combining massive geometry with deconstructivism.	2, 3, 1, 0, 0, 0, 0	38	1, 3, 1, 0, 0, 0, 0	32
6	S2–P3	A desert-like restaurant on a hillside with a desert-looking landscape, geometrical shape with minimalist form.	2, 2, 2, 1, 1, 0, 0	45	2, 2, 1, 1, 2, 0, 0	43
7	S2–P4	A desert-inspired restaurant rising above an 8-m slope, featuring organic floating terraces, curved futuristic forms, and integrated water features.	3, 2, 2, 1, 1, 1, 0	55	3, 2, 1, 2, 2, 2, 0	62
8	S3–P1	Floating nature restaurant, wooden decks, open-air, surrounded by trees and a river.	2, 1, 2, 1, 1, 1, 0	42	2, 1, 3, 2, 1, 1, 0	52
9	S3–P2	River sky lounge stepped terrace overlooking river, modern and elegant form.	2, 2, 2, 1, 0, 1, 0	45	2, 2, 2, 1, 0, 2, 0	48
10	S3–P3	River life restaurant, inspired by fishing culture.	1, 1, 1, 0, 0, 0, 0	18	1, 1, 1, 1, 0, 1, 0	27
11	S4–P1	A modern Möbius strip with sharp edges and glass surfaces inspired by the concept of oblivion.	2, 3, 1, 1, 1, 1, 0	50	2, 3, 1, 2, 0, 1, 0	52
12	S4–P2	Oblivion sharp irregular sides with triangular top form restaurant.	2, 2, 1, 0, 0, 0, 0	32	2, 2, 2, 0, 0, 0, 0	37
13	S4–P3	Futuristic restaurant inspired by the concept of a parallel universe.	2, 2, 0, 0, 0, 0, 0	27	2, 2, 0, 1, 0, 0, 0	32
14	S4–P4	A simple layout restaurant that reflects and shows the memories of the customers.	1, 1, 0, 0, 0, 0, 0	13	0, 0, 0, 1, 0, 0, 0	5
15	S5–P1	A riverside restaurant inspired by sand and Islamic architecture, featuring earthy tones and arched doorways, blending tradition with natural beauty.	2, 3, 1, 1, 2, 0, 0	50	2, 3, 1, 1, 3, 0, 0	53
16	S5–P2	A luxurious riverside restaurant with a culturally inspired design blending Moroccan arches, Japanese minimalism, Mediterranean terraces, and African textures.	3, 3, 2, 1, 2, 1, 0	65	3, 3, 2, 0, 2, 0, 0	57

Table 5. *Cont.*

No.	Student	Prompt Text	Rater-1 Scores		Rater-2 Scores	
			S1, S2, S3, S4, S5, S6, S7	R	S1, S2, S3, S4, S5, S6, S7	R
17	S6–P1	A restaurant on a bridge or raised platform with transparent glass walls and sleek structure, looking light and floating, giving wide views of the landscape.	2, 2, 2, 1, 0, 1, 0	45	1, 2, 2, 1, 0, 1, 0	38
18	S6–P2	A restaurant near a waterfall with natural shapes, sloped roofs, and big glass walls. The design connects the building with the sound and water movement.	2, 2, 2, 2, 1, 1, 0	53	3, 2, 2, 2, 2, 1, 0	63
19	S6–P3	Restaurant in the desert with simple, open-air architecture and clean lines. Use natural materials and shaded terraces to match the landscape.	2, 2, 2, 1, 1, 1, 0	48	2, 2, 1, 1, 1, 1, 0	43
20	S7–P1	A design inspired by travel, a place for tourists, a place like a resort.	1, 1, 0, 0, 0, 0, 0	13	1, 1, 1, 0, 0, 0, 0	18
21	S7–P2	A warm, earthy design inspired by the cycle of nature, from soil to plate.	1, 1, 0, 1, 2, 0, 0	25	1, 1, 0, 0, 2, 0, 0	20
22	S7–P3	A modern, sophisticated restaurant set in the city, inspired by light, water, and glass.	2, 2, 0, 2, 1, 1, 0	43	2, 2, 0, 2, 1, 1, 0	43
23	S9–P1	A restaurant like Babylon times.	1, 1, 0, 0, 0, 0, 0	13	1, 1, 1, 0, 0, 0, 0	18
24	S9–P2	A restaurant floated between timelines.	1, 1, 0, 0, 0, 0, 0	13	1, 2, 0, 0, 0, 0, 0	20
25	S9–P3	A restaurant inspired by ancient Silla myth.	1, 1, 0, 0, 0, 0, 0	13	1, 1, 0, 1, 0, 0, 0	18
26	S10–P1	A bridge that connects two riverbanks invites you to dine above the water. Smooth, flowing forms that curve and twist to create cosy dining pods and open lounges, with metallic tones.	3, 2, 2, 1, 2, 1, 0	58	3, 2, 2, 1, 1, 1, 0	55
27	S10–P2	A minimalist circular restaurant floating on a calm river, connected by a slender wooden walkway. Smooth matte white curved roof, full-height glass walls, light natural wood interior.	3, 3, 3, 1, 3, 2, 0	77	3, 3, 3, 2, 3, 2, 0	82
28	S11–P1	Restaurant carved into rock formation beside river, stone and concrete textures, terraces.	2, 2, 2, 1, 2, 1, 0	52	2, 1, 1, 3, 2, 2, 0	53
29	S11–P2	Floating modular restaurant on a calm river, glass and metal structure, reflecting water surface, dynamic architectural form.	3, 2, 2, 2, 1, 2, 0	63	3, 2, 2, 1, 1, 2, 0	58
30	S12–P1	A high-altitude remote location, a leftover of an old telescope array with massive parabolic dishes.	2, 2, 2, 1, 1, 1, 0	48	1, 2, 2, 1, 2, 0, 0	42
31	S12–P2	A restaurant that is built around a meteor crash site, blended with futuristic design.	2, 2, 1, 1, 1, 1, 0	43	2, 2, 1, 1, 2, 1, 0	47
32	S12–P3	A mountain with a deep slope, with fractured rocks and displaced soil layers, a restaurant built among the remains with a view of a stream below.	3, 2, 3, 2, 2, 1, 0	68	3, 2, 3, 2, 2, 0, 0	65
33	S13–P1	A sci-fantasy restaurant constructed from stacked vintage CRT televisions.	2, 3, 1, 1, 1, 1, 0	50	2, 3, 1, 1, 1, 0, 0	47

Table 5. *Cont.*

No.	Student	Prompt Text	Rater-1 Scores		Rater-2 Scores	
			S1, S2, S3, S4, S5, S6, S7	R	S1, S2, S3, S4, S5, S6, S7	R
34	S13–P2	A restaurant caught in a temporal distortion, architecture warped and twisted with impossible angles, walls and roof stretched and rotated.	2, 3, 2, 1, 0, 2, 0	55	2, 3, 2, 2, 0, 1, 0	57
35	S14–P1	Glowing spherical dining room that appears to float above the site.	2, 2, 1, 2, 0, 1, 0	45	2, 2, 1, 2, 0, 1, 0	45
36	S14–P2	Trees and biophilic design, shelter, growth, balance restaurant.	1, 1, 1, 0, 0, 0, 0	18	1, 1, 1, 0, 0, 0, 0	18
37	S15–P1	Floating crystal restaurant above a calm river, inspired by Aurora lights. Geometric design.	2, 3, 2, 2, 2, 1, 0	63	2, 3, 2, 2, 2, 1, 0	63
38	S15–P2	Pyramid-shaped restaurant blending ancient Egyptian architecture with subtle steampunk elements. Elegant futuristic fantasy design with golden lighting, metal details, and soft atmosphere.	3, 3, 1, 3, 2, 2, 0	73	3, 3, 1, 3, 2, 1, 0	70
39	S16–P1	Restaurant that physically transforms with the seasons, featuring dynamic, adaptive exterior panels and change.	2, 2, 1, 1, 0, 1, 0	40	2, 2, 1, 1, 1, 1, 0	43
40	S17–P1	A large, modern cargo ship grounded in a post-apocalyptic river that has been turned into a restaurant.	3, 2, 2, 1, 1, 1, 0	55	3, 3, 1, 1, 0, 1, 0	53
41	S17–P2	A futuristic-looking restaurant in a post-apocalyptic world; the building is on a slope next to a river.	2, 2, 2, 1, 0, 1, 0	45	2, 2, 2, 2, 0, 1, 0	50
42	S17–P3	A restaurant built inside a tree, more like a treehouse, located next to a river on a sloped site.	2, 1, 2, 1, 1, 1, 0	42	2, 1, 3, 3, 1, 3, 0	63
43	S18–P1	A restaurant inspired by reflection, both physical (mirrors, water) and emotional (self-awareness).	2, 2, 0, 1, 1, 0, 0	35	2, 2, 0, 0, 1, 1, 0	33
44	S18–P2	A restaurant inspired by flowing water. Its shape should mimic waves and currents. The structure should feel fluid and dynamic.	2, 2, 1, 1, 1, 1, 0	43	2, 2, 0, 1, 1, 1, 0	38
45	S19–P1	Aqueous zenith, futuristic restaurant inspired by the flow of water and wind and the movement of time.	2, 2, 0, 1, 1, 1, 0	38	2, 2, 0, 1, 1, 2, 0	42
46	S19–P2	The garden mind, nature mirrors human's mind in every space, representing stages of emotional growth.	2, 2, 1, 1, 1, 0, 0	40	1, 2, 0, 1, 1, 0, 0	28
47	S19–P3	Floating bubble, a transparent spherical restaurant that dramatically floats above the river, offering panoramic views and magical dining.	3, 2, 2, 1, 1, 2, 0	58	2, 2, 3, 1, 0, 0, 0	47
48	S20–P1	A restaurant with an army look, showing strong military elements, metal structure with bold geometrical form.	2, 2, 1, 0, 1, 1, 0	38	2, 2, 0, 1, 1, 1, 0	38
49	S20–P2	A restaurant inspired by Raven, having a dark tone, deep purple and silver. With features resembling mystery-like wings.	2, 2, 1, 1, 3, 1, 0	50	2, 2, 1, 0, 3, 0, 0	42

Table 5. *Cont.*

No.	Student	Prompt Text	Rater-1 Scores		Rater-2 Scores	
			S1, S2, S3, S4, S5, S6, S7	R	S1, S2, S3, S4, S5, S6, S7	R
50	S21–P1	Design a brutalist futurist riverside restaurant on a slope.	2, 3, 2, 0, 0, 0, 0	43	2, 3, 3, 1, 1, 0, 0	57
51	S21–P2	Design a volcanic-erupted riverside restaurant on a slope, with lava as the construction material.	2, 2, 2, 1, 2, 1, 0	52	2, 2, 2, 0, 2, 0, 0	43
52	S22–P1	A futuristic multi-story luxury restaurant building situated by a river, inspired by the elegant form of a chef's hat.	3, 2, 1, 1, 0, 2, 0	50	3, 2, 0, 3, 0, 3, 0	58
53	S22–P2	A restaurant by the riverside inspired by a coffee mug mixed with curved lines and organic form.	2, 2, 1, 0, 0, 1, 0	35	2, 2, 0, 0, 0, 2, 0	33
54	S23–P1	A restaurant on the river shores; the land slope varies from 2 to 8 m, getting inspiration from water and the river.	2, 1, 2, 1, 1, 0, 0	38	3, 1, 2, 1, 2, 0, 0	48
55	S23–P2	A two-story restaurant on the river shore using organic shapes, local cultural environment and materials.	2, 2, 2, 1, 2, 1, 0	52	3, 3, 2, 0, 1, 0, 0	53

The dataset reveals substantial within-student variance in several cases. S2's four prompts range from R = 38 (P2: [A futuristic restaurant inspired by ancient Mesopotamia, combining massive geometry with deconstructivism]) to R = 62 (P1: full slope, geometry, and style specification). S4's four prompts range from R = 13 (P4: purely emotional descriptor) to R = 50 (P1: Möbius strip with glass surfaces). This variance indicates that prompt quality within a single student's submission is inconsistent, suggesting that students have an intuitive sense of when to add detail but do not apply it systematically.

4.2. Per-Student Summary and Composite Scores

Table 6 aggregates scores per student, showing average dimension scores, composite average R, best single-prompt R, and overall richness tier based on the average R.

Table 6. Average Dimension Scores, Composite R, Best Prompt R, and Richness Tier.

ID	n	S1	S2	S3	S4	S5	S6	S7	Avg R	Best R	Tier
S1	3	2	2	1	1	1	1	0	43	43	Moderate
S2	4	2.5	2.5	1.75	0.75	0.75	0.5	0	50	62	Moderate
S3	3	1.67	1.33	1.67	0.67	0.33	0.67	0	35	45	Moderate
S4	4	1.75	2	0.5	0.25	0.25	0.25	0	31	50	Moderate
S5	2	2.5	3	1.5	1	2	0.5	0	58	65	Moderate
S6	3	2	2	2	1.33	0.67	1	0	49	53	Moderate
S7	3	1.33	1.33	0	1	1	0.33	0	27	43	Sparse
S9	3	1	1	0	0	0	0	0	13	13	Sparse
S10	2	3	2.5	2.5	1	2.5	1.5	0	68	77	Rich
S11	2	2.5	2	2	1.5	1.5	1.5	0	58	63	Moderate
S12	3	2.33	2	2	1.33	1.33	1	0	53	68	Moderate
S13	2	2	3	1.5	1	0.5	1.5	0	53	55	Moderate

Table 6. *Cont.*

ID	n	S1	S2	S3	S4	S5	S6	S7	Avg R	Best R	Tier
S14	2	1.5	1.5	1	1	0	0.5	0	32	45	Moderate
S15	2	2.5	3	1.5	2.5	2	1.5	0	68	73	Rich
S16	1	2	2	1	1	0	1	0	40	40	Moderate
S17	3	2.33	1.67	2	1	0.67	1	0	47	55	Moderate
S18	2	2	2	0.5	1	1	0.5	0	39	43	Moderate
S19	3	2.33	2	1	1	1	1	0	45	58	Moderate
S20	2	2	2	1	0.5	2	1	0	44	50	Moderate
S21	2	2	2.5	2	0.5	1	0.5	0	48	52	Moderate
S22	2	2.5	2	1	0.5	0	1.5	0	43	50	Moderate
S23	2	2	1.5	2	1	1.5	0.5	0	45	52	Moderate

The average R across the 22 analysed students was 45. Most students (18 of 22) fell in the Moderate tier. Two students, S10 (avg R = 68) and S15 (avg R = 68), achieved Rich-tier averages. Two students, S7 (avg R = 27) and S9 (avg R = 13), fell below the Sparse threshold (R < 30); two further students, S4 (avg R = 31) and S14 (avg R = 32), scored just above that threshold in the low-Moderate range but showed similarly limited dimensional coverage. No student achieved a Highly Specific average (R $\geq$ 80), though S10–P2 reached the highest single-prompt score of R = 77. Under Rater 2, the same prompt (S10–P2) scores R = 82 (Highly Specific tier), reflecting disagreement on the lighting and quality-modifier dimensions for this prompt; this does not alter the primary Rater-1-based classification used throughout the paper but is noted here for transparency.

4.3. Dimension-Level Class Analysis

The class-wide performance across the seven dimensions (measured by average score, percentage of the maximum achieved, gap to the dimensional ceiling, and improvement-priority classification) highlights the extent to which students depend on each dimension (Table 7).

Table 7. Class-Wide Dimension Performance, Gap to Ceiling, and Improvement Priority.

Dimension	Weight	Max	Class Avg	% of Max	Gap to Ceiling	Improvement Priority
Subject specificity	2	3	2.08	69%	0.92	Low-well covered
Style & medium	2	3	2.04	68%	0.96	Low-well covered
Composition & framing	1.5	3	1.34	45%	1.66	Medium-partial adoption
Lighting & atmosphere	1.5	3	0.95	32%	2.05	High-critical gap
Color & palette	1	3	0.95	32%	2.05	High-critical gap
Quality modifiers	1	3	0.85	28%	2.15	High-critical gap
Negative clauses	1	2	0	0%	2.00	High-critical gap

Subject specificity (S_1) and style/medium (S_2) are the two dimensions students engage with most naturally, achieving 69% and 68% of their respective maxima. These dimensions map onto concepts "what is the building?" and "what does it look like stylistically?" that architecture students are trained to articulate through other channels (design briefs, precedent studies, mood boards).

Composition and framing (S_3) show intermediate adoption at 45% of maximum. Some students intuitively include spatial cues (overlooking the river, stepped terrace), but few use camera-angle or shot-type language that would more precisely direct the generator's spatial framing.

Lighting/atmosphere (S_4), colour/palette (S_5), and quality modifiers (S_6) all fall in the High-priority gap zone, each achieving under 33% of maximum. These represent the class's biggest opportunity for rapid improvement: each of these dimensions can be addressed with a single targeted phrase appended to any existing prompt without restructuring the core concept.

Negative clauses (S_7) achieved a class mean of exactly 0.00; not a single student in the cohort used a negative clause or exclusion term in any of their 55 prompts. This complete absence is the most actionable finding of the experiment, since negative clauses are among the most effective mechanisms for suppressing common generator artefacts (generic people, text overlays, flat backgrounds, photorealistic clutter) that degrade architectural image quality.

4.4. Exemplary and Weakest Prompts

Table 8 presents the seven most instructive prompts from the dataset, the four highest scoring (all Rich-tier) and the three lowest scoring (all Sparse-tier), with annotations explaining what drives each score (see Appendix A).

Table 8. Exemplary (Rich-Tier) and Weakest (Sparse-Tier) Prompts from the Dataset.

ID	Category	R	Prompt Text	Why Does it Score That Way
S10–P2	Highest R	77	A minimalist circular restaurant floating on a calm river, connected by a slender wooden walkway. Smooth matte white curved roof, full-height glass walls, light natural wood interior.	Maximum S1 + S3 + S5; specific material palette; precise spatial framing; style coherent.
S15–P2	Rich style	73	Pyramid-shaped restaurant blending ancient Egyptian architecture with subtle steampunk elements. Elegant futuristic fantasy design with golden lighting, metal details, and soft atmosphere.	Only prompt scoring 3/3 on both S2 and S4 simultaneously; layered atmosphere and named material detail.
S5–P2	Rich multicultural	65	A luxurious riverside restaurant with a culturally inspired design blending Moroccan arches, Japanese minimalism, Mediterranean terraces, and African textures.	Four named cultural references create precise style-token density; the subject is fully described.
S12–P3	Rich site	68	A mountain with a deep slope, with fractured rocks and displaced soil layers, a restaurant built among the remains with a view of a stream below.	Maximum S3; site topography precisely described; spatial sequence (slope, rocks, stream view) grounds the prompt.
S9–P1	Sparse concept only	13	A restaurant like Babylon times.	Single simile; zero spatial, material, compositional, or lighting information; generator left to fill all dimensions.

Table 8. *Cont.*

ID	Category	R	Prompt Text	Why Does it Score That Way
S7–P1	Sparse vague intent	13	A design inspired by travel, a place for tourists, a place like a resort.	Three repeated near-synonymous intent tokens with no architectural, material, or visual information.
S4–P4	Sparse lowest R	13	A simple layout restaurant that reflects and shows the memories of the customers.	Lowest R in dataset; purely emotional/programmatic statement; no spatial, style, colour, or lighting cues at all.

4.5. Prompt-Output Richness Correlation

Students are ranked by descending average R. For each student, average R, best single-prompt R, Output O score, absolute gap |R-O|, both tier classifications, and a description of the AI-generated outputs are provided (Table 9).

Table 9. Prompt Richness vs. Output Richness, Full Correlation Dataset, Ranked by Average R.

Student	Prompts	Avg R	Best R	Output O	$\lvert R-O \rvert$	Prompt Tier	Output Tier	Output Description
S10	2	68	77	68	0	Rich	Rich	Best class output: precise circular geometry, reflective water surface, accurate matte-white material, calm atmospheric quality
S15	2	68	73	65	3	Rich	Rich	Pyramid (P2) most thematic output in class; golden tones faithfully reproduced; aurora crystal (P1) vibrant but compositionally loose
S5	2	58	65	62	4	Moderate	Rich	Strong ornamental detail; warm, earthy tones; arched forms faithfully reproduced; multicultural layering visible in P2 output
S11	2	58	63	52	6	Moderate	Moderate	Stone textures and reflective metal are faithfully rendered; the P2 floating structure shows dynamic form; both outputs are above average.
S12	3	53	68	44	9	Moderate	Moderate	P3 rocky topography most site-specific; telescope array (P1) imaginative but underlit; meteor crash (P2) generic futurism
S13	2	53	55	50	3	Moderate	Moderate	CRT-TV stacking (P1) rendered distinctively; temporal distortion (P2) achieves warped geometry; high conceptual fidelity

Table 9. *Cont.*

Student	Prompts	Avg R	Best R	Output O	|R−O|	Prompt Tier	Output Tier	Output Description
S2	4	50	62	58	8	Moderate	Moderate	Futuristic layered terracing with river context; material richness above class average; crystal geometry partially resolved
S6	3	49	53	48	1	Moderate	Moderate	Glass-heavy structures rendered cleanly; waterfall P2 result most atmospheric; desert P3 competent but generic
S21	2	48	52	43	5	Moderate	Moderate	Brutalist massing (P1) rendered with confidence despite sparse prompt; volcanic lava (P2) strong drama; model priors evident
S17	3	47	55	40	7	Moderate	Moderate	Cargo ship (P1) concept recognisable; post-apocalyptic mood weak; treehouse (P3) charming but style tag generic
S19	3	45	58	42	3	Moderate	Moderate	Floating bubble (P3) distinctive and correctly spherical; garden mind (P2) abstract and hard to read; P1 flow concept rendered competently
S23	2	45	52	40	5	Moderate	Moderate	Organic riverside form rendered moderately; local material cues partially conveyed in P2; slope topography visible in both outputs
S20	2	44	50	38	6	Moderate	Moderate	Raven prompt (P2) dark tonal output with wing-like cantilevers; military (P1) render lacks material depth and atmosphere.
S1	3	43	43	42	1	Moderate	Moderate	Organic fluid form rendered with wave-like sweeps; moderate material detail; lighting flat; style partially captured
S22	2	43	50	32	11	Moderate	Moderate	Chef-hat form (P1) recognisable in spiral tower; coffee mug (P2) organic but flat; outputs lack lighting and material specificity
S16	1	40	40	30	10	Moderate	Moderate	Adaptive facade rendered as static wavy form; seasonal transformation concept not visually apparent; single output limits assessment

Table 9. *Cont.*

Student	Prompts	Avg R	Best R	Output O	|R−O|	Prompt Tier	Output Tier	Output Description
S18	2	39	43	33	6	Moderate	Moderate	Fluid waveforms are produced, but generically; mirror/water reflective theme is partially conveyed in P1; material specificity is absent.
S3	3	35	45	32	3	Moderate	Moderate	Open-air wooden deck generically rendered; atmosphere weak; fishing culture concept lost in P3 output.
S14	2	32	45	24	8	Moderate	Sparse	Glowing sphere (P1) recognisable but generically lit; biophilic P2 produced a cliché treehouse; low material specificity throughout
S4	4	31	50	25	6	Moderate	Sparse	Abstract futuristic shapes poorly grounded; spatial logic unclear; Möbius concept recognisable in P1 but flat lighting throughout
S7	3	27	43	20	7	Sparse	Sparse	Generic modern restaurant outputs; no distinctive style; bland neutral lighting; conceptual intent absent from all three outputs
S9	3	13	13	18	5	Sparse	Sparse	Vague ancient-themed renders; low detail; inconsistent style across three outputs; compositionally undifferentiated

Pearson's r between per-student average R (aggregated across each student's 1–4 submitted prompts) and portfolio-level O across the 22 analysed students was r = 0.940, 95% CI [0.86, 0.98] (Figure 2).

This represents a near-perfect positive linear relationship: knowing a student's prompt richness score predicts their output richness score with approximately 88% of variance explained ($R^2 \approx 0.88$). The class average output O was 41 against the average prompt R of 45 (Table 10), a difference of 4 points, a small positive residual suggesting that AI generators marginally supplement underdeveloped prompts using training-data priors, but this effect is minor relative to the variance explained by prompt quality itself.

Table 10. Descriptive Statistics for Prompt Richness (R) and Output Richness (O), n = 22.

Measure	Mean	SD	Minimum	Maximum	n
Average R (per student)	44.95 (≈45)	12.84	13 (S9)	68 (S10, S15)	22
Best-prompt R (per student)	52.50	13.42	13 (S9)	77 (S10)	22
O (per student)	41.18 (≈41)	14.10	18 (S9)	68 (S10)	22

Three structural patterns were identified in the correlation data:

- Rich-tier convergence (O almost equal to R): S10 (R = 68, O = 68, gap = 0) and S15 (R = 68, O = 65, gap = 3) show near-perfect prompt-to-output translation. These

students specified subject, style, composition, colour, and, in the case of S15, lighting, covering five of seven dimensions at or above average, leaving the generator with minimal interpretive freedom and correspondingly precise outputs.

- Positive divergence (O > R): S5 (R = 58, O = 62), S2 (R = 50, O = 58), and S21 (R = 48, O = 43) each received outputs richer than their prompt scores would predict. In S5's case, named cultural references (Moroccan, Japanese, Mediterranean, African) appear to function as high-density style tokens (terms likely to be well-represented in the training corpora of major diffusion models), enabling the generator to infer extensive formal and material detail from relatively compact prompt text [35]. In S21's case, the term "brutalist futurist" may similarly activate established visual priors for massing, material, and texture; however, as with the S5 case above, this interpretation is speculative and requires controlled ablation studies across multiple generator platforms to confirm.
- Sparse clustering: S7 (R = 27, O = 20), S9 (R = 13, O = 18), and S4 (R = 31, O = 25) cluster at the bottom of both scales. Their prompts provided insufficient information for the generator to resolve architectural form, material, or atmosphere, resulting in outputs that are generic, stylistically inconsistent, and architecturally non-specific. The slight positive divergence in S9 (O = 18 vs. R = 13) reflects a floor effect: even the most minimal prompt triggers some architectural imagery, preventing outputs from scoring below approximately 15.

Because O's seven dimensions were deliberately mirrored to R's structure (Section 3.5.1), part of the strength of this correlation is attributable to shared design rather than fully independent measurement. As a robustness check, the same three judges additionally provided a single holistic quality rating (1–10) per student, assigned independently of the seven O dimensions and without reference to R (Table 11).

Table 11. Correlation Between R and an Independent Holistic Quality Rating.

Measure Pair	Pearson r	95% CI	*p*-Value	n
Average R vs. composite O (primary analysis)	0.940	[0.86, 0.98]	<0.001	22
Average R vs. independent holistic rating (1–10)	0.851	[0.67, 0.94]	<0.001	22

5. Discussion

5.1. Prompt Richness as Architectural Language

The correlation result (r = 0.940) carries a clear disciplinary implication: writing an effective AI prompt for architectural visualisation is structurally equivalent to writing a design brief. Both require the author to spatially ground a concept (subject specificity, composition), give it a visual register (style, medium), specify its atmospheric qualities (lighting, colour), and define its boundaries (negative clauses, exclusions). Students who addressed all seven dimensions (whether they had prior familiarity with the rubric) consistently produced architecturally richer outputs.

This reframes prompt engineering not as a peripheral technical skill within design thinking, but as a direct expression of architectural intelligence, specifically its verbal dimension. As documented by Avidan and Goldschmidt [27] and Cikis and Ipek Ek [22], students who can articulate spatial and material qualities with clarity tend to produce more developed and refined design representations. This positions verbal specification as one input alongside, rather than a replacement for, the sketching, reflection, and iterative reasoning that remain central to design pedagogy [22,25,27]. Framed this way, the PRI

complements rather than constrains established studio practices such as crit sessions, portfolio review, and instructor critique, which continue to provide the qualitative, reflective assessment that a rubric of this kind cannot capture. The PRI rubric makes this verbal-spatial translation explicit and measurable. A student who can precisely describe a building's form, material, light, and spatial experience in written language is performing the same cognitive task as a student who can precisely sketch it. The rubric in this paper provides a structured pathway for developing that verbal and spatial translation capacity.

5.2. Generator Effects and Their Pedagogical Implications

The positive divergence cases (S5, S21) reveal an important nuance: AI generators do not passively transcribe prompts. They actively complete underspecified prompts using statistical patterns from their training data. A prompt that invokes 'brutalist' architecture or 'Moroccan arches' is not only a form; it is a pointer into a dense cluster of training examples, each carrying implied formal, material, and atmospheric qualities that the generator applies even without explicit instruction.

This has a dual pedagogical implication. On one hand, students who use culturally or historically specific style tokens can achieve richer outputs from shorter prompts than the rubric score alone might predict. This constitutes a pedagogically transferable strategy: identifying style tokens that are densely represented in generator training data and that carry substantial implied formal and atmospheric properties.

On the other hand, over-reliance on style tokens without spatial or atmospheric grounding produces outputs that are stylistically recognisable but architecturally imprecise; the generator fills in a generic form of the named style, not a site-specific or concept-specific interpretation. The rubric's emphasis on compositional, atmospheric, and chromatic specification (S_3–S_5) serves to counteract over-reliance on style tokens by requiring explicit spatial and sensory grounding. The style-token density hypothesis (that culturally or historically specific terms activate particularly rich clusters of training examples in diffusion models) cannot be confirmed without controlled ablation studies that systematically vary style tokens while holding other prompt dimensions constant, and across multiple generator platforms (e.g., Midjourney, DALL-E 3, Stable Diffusion, Leonardo AI). Such studies represent a clear direction for future work and would provide empirical evidence to refine both the weighting scheme and the pedagogical recommendations. In the absence of such ablation data, instructors should guide students to treat style tokens as a complement to, rather than a substitute for, explicit spatial, material, and atmospheric specification.

5.3. The Negative Clause Finding

The complete absence of negative clauses across all prompts (S_7 mean = 0.00) is the most striking and practically actionable finding in the dataset. Negative clauses (instructions telling the generator what to exclude or avoid) are among the most effective mechanisms available to users of text-to-image systems. They can suppress text and watermarks that render architectural images unpublishable; generic photorealistic people that distract from architectural form; flat or overexposed backgrounds that reduce spatial depth; cartoon or illustration styles that conflict with a professional architectural register; and specific visual artefacts common to a particular generator.

The complete non-adoption of this mechanism by all students strongly suggests that it was unknown to them, not that they consciously chose not to use it. Because $S_7 = 0$ for every one of the 55 prompts in this dataset, this dimension has zero variance and its relationship to output richness (O) could not be empirically tested here; the practical importance we attribute to negative clauses rests on prior prompt-engineering literature [37]

rather than on evidence internal to this study. This finding identifies the highest-leverage single instructional intervention available within the studio context. A brief targeted demonstration using paired image comparisons (with and without negative clauses) is predicted to increase S_7 scores from zero to at least 1 for most students.

5.4. Within Student Variance

The wide variance observed in students like S2 and S4 indicates that students are not operating at a fixed prompt quality level. They are making ad hoc decisions about how much detail to include, likely influenced by how confident they feel about a particular concept. Concepts that students could readily verbalise generated richer prompts; abstract or emotionally framed concepts, which resist direct verbal specification, generated sparser ones.

This pattern suggests that the diagnostic-iterative loop (described in Step 10 of the procedure) would be particularly effective for these students. Their highest-scoring prompts demonstrate that they are capable of Rich-tier writing; the challenge is applying that capability consistently across all prompts, regardless of concept type.

5.5. Practical Implications for Architectural Pedagogy

The correlation reported in Section 4.5 (r = 0.940) demonstrates that prompt richness, as operationalised by R, statistically predicts output richness, as operationalised by O. This is a technical, measurement-level finding. It does not, by itself, establish that teaching students to maximise their R score improves broader design-pedagogical outcomes such as critical thinking, reflexivity, or individual creative expression-claims that would require a separate, controlled pedagogical study (e.g., comparing design outcomes, conceptual diversity, and reflective capacity between cohorts taught with and without the PRI rubric). We caution against conflating the demonstrated statistical association with pedagogical validity, and we discuss safeguards against over-standardisation in Section 6.

Whether the verbal-specification skills developed through the PRI transfer to downstream technical workflows such as BIM parameter-setting or quantitative estimation is an open question that this study does not test; we flag it as a direction for future research rather than a demonstrated outcome.

6. Conclusions

This paper has introduced a proposed ten-step numerical procedure for evaluating the richness of text-to-image AI prompts, applied in full to a dataset of 55 prompts produced by 22 architecture students in a design studio context. The procedure generates a composite Prompt Richness Index (R) from seven weighted dimensions, and a parallel Output Richness Index (O) applied to the AI-generated images produced from those prompts. Pearson's r between R and O yielded r = 0.940 ($p < 0.001$), establishing that prompt richness is the primary determinant of output richness in this context.

Three findings stand out as particularly significant. First, no student achieved a Highly Specific prompt average ($R \geq 80$), and the class mean of R = 45 sits at the middle of the Moderate tier (indicating that a substantial performance ceiling remains available to every student in this cohort). Second, the complete non-adoption of negative clauses across all prompts represents a missed opportunity that requires a single targeted pedagogical intervention to address. Third, the generator prior effect (where culturally specific style tokens yield outputs richer than their R scores predict) identifies a learnable strategy that complements dimensional expansion in prompt development.

The prompt writing for AI visualisation is not a technical add-on to design thinking. It is a new form of spatial literacy (the ability to translate an architectural concept into precise,

layered, atmospheric language that a generative system can render with fidelity). This paper provides both an empirical foundation for that claim and a practical, preliminary instrument for developing this competency in design students, pending further psychometric validation and replication across design tasks, student populations, and institutional contexts. As text-to-image tools become standard in architectural practice and education, prompt literacy belongs in the design studio curriculum together with drawing, modelling, and representation.

Based on the findings, eight recommendations are proposed for integrating prompt literacy instruction into architecture design studios that use AI visualisation tools:

1. Introducing Prompt Anatomy Early: Before students use AI tools for the first time in a project, the seven-dimensional rubric should be presented as a prewriting checklist. Students annotate their draft prompt against the seven dimensions, identify any zeros, and add at least one descriptor per missing dimension before generating. This converts the rubric from an assessment instrument into a design thinking framework.
2. Lighting Variable Exercise: The S_4 (lighting/atmosphere) gap is the easiest to address with a direct demonstration. Students write a single prompt, generate an image, then add one lighting descriptor (warm golden-hour side light, cold grey overcast diffuse) and generate again. Comparing the two outputs side-by-side makes the impact of S_4 immediately visible and builds an intuitive vocabulary for atmospheric specification.
3. Negative Clause Workshop: A dedicated 20-min workshop demonstrates negative clauses through before/after image pairs. Show outputs with and without (no text, no people, avoid photorealistic style). The contrast between outputs with and without negative clauses is sufficiently pronounced to make the mechanism self-evident to students, reducing the need for extended theoretical explanation. Students then apply at least two targeted negative clauses to their next AI submission.
4. Prompt Peer-Review: Before any AI generation session, students exchange prompts and apply the rubric to each other's work. The peer review identifies the single lowest-scoring dimension and suggests one specific phrase to improve it. This builds rubric fluency, encourages reflective writing, and generates a larger volume of feedback than instructor-only review can provide.
5. Iterative Re-Scoring Loop: The studio brief should require students to submit a minimum of two prompt iterations per design stage, with R scores and dimension annotations for each. The first prompt establishes a baseline; the second incorporates targeted improvements to the lowest-scoring dimension. Both prompt iterations and the corresponding generated images are submitted together, accompanied by a brief written reflection addressing the specific changes made and their justification.
6. Cross Tool Comparison: Assigning the same Rich-tier prompt to two different generators (e.g., Midjourney and DALL-E 3) reveals how generator training-data priors interact with prompt specificity. Students observe that identical prompt text can yield stylistically different outputs, deepening their understanding of the distinction between prompt content (what they control) and generator character (what they do not). This also builds independent platform prompt literacy.
7. End of Semester Prompt Portfolio: A prompt portfolio submitted at the end of the design studio (comprising all prompts used across the project, with R scores and reflective annotations) provides summative evidence of growth in prompt literacy. Marking criteria should credit demonstrated engagement with the iterative revision process (e.g., evidence of diagnosis and targeted revision between prompt iterations) rather than treating R-score improvement as the primary or sole basis for grading; quantitative R trends should be considered alongside, and never in place of, qualitative

instructor assessment of each student's conceptual development, reflective reasoning, and individual design voice.

8. Safeguarding Creative Agency: Because rubric-based instruction carries an inherent risk of encouraging formulaic or convergent prompting, instructors should explicitly frame the PRI as a formative, pre-generation self-check rather than a scoring target to be maximised. Prompt peer review (Recommendation 4) should evaluate distinctiveness and individual voice alongside rubric completeness, and instructors should periodically showcase high-R prompts that nonetheless differ substantially from one another in concept and language, to demonstrate that dimensional richness and creative diversity are compatible rather than competing goals.

To bridge theoretical instruction with practical execution, the proposed 7-dimension Prompt Richness Index (PRI) is structured to support the tripartite mechanism of architectural education: theoretical modules, technical workshops, and design studios. In early-semester theoretical courses, prompt dimensions such as composition and framing (D3) and lighting and atmosphere (D4) reinforce historical and typological discourse. In mid-semester technical workshops, rendering detail and technical quality (D6) and colour and palette (D5) anchor generative exploration in constructability. Ultimately, in advanced design studios, the synthesised application of all seven dimensions enables students to articulate complex programmatic narratives across multi-story plans, sections, and 3D perspectives.

Several limitations should be noted. First, the study is based on a single cohort (n = 22) and a single design task. Results may not generalise to other student populations, design briefs, or AI platforms. Second, the study does not control students' prior experience with AI tools, which may confound prompt quality independently of design ability. Third, the single O score assigned per student aggregates across multiple images, which may mask within-student output variance. Future studies should assign O at the individual prompt-image level to enable within-student correlation analysis. Fourth, R and O share a mirrored dimensional structure (Section 3.5.1), meaning part of their correlation reflects shared design rather than fully independent measurement. Fifth, students used a variety of unspecified, self-selected text-to-image generators, introducing an uncontrolled platform-level variable. Sixth, this pilot study involved a single design brief and a single institution; applicability to other design tasks, student levels, and educational contexts remains to be tested. Seventh, the Prompt Richness Index was developed specifically for text-to-image generation in the context of architectural concept visualisation. Eighth, the author also served as Rater 1 for the primary Prompt Richness analysis; while a second, independent rater was used to assess reliability, the primary results were not scored by a fully independent party, which is a limitation on the objectivity of the R dataset. The framework's applicability to other AI-assisted design activities such as concept development using large language models, image-to-image editing, three-dimensional generative modelling, or building information modelling (BIM) workflows remains untested and is an important direction for future research. The relationship between prompt richness and design outcomes in these other modalities may differ substantially, and the weights and dimensions proposed here may not transfer directly. Researchers and educators should therefore exercise caution when generalising the framework beyond the specific task and context examined in this study.

Funding: This research received no external funding.

Institutional Review Board Statement: Tishk International University does not currently maintain a specific written institutional regulation requiring formal Institutional Review Board (IRB) approval for research based solely on anonymised routine educational activities conducted within the normal

teaching and assessment process. Studies of this nature are not considered human-subject research requiring formal ethical review by an IRB or equivalent ethics committee. Therefore, the study was approved for exemption by the Dean of Students at Tishk International University.

Informed Consent Statement: Because the study relied exclusively on anonymised educational records generated through routine coursework and involved no additional interaction, intervention, or collection of identifiable personal information, separate written informed consent was waived. The exemption was approved by the Dean of Students at Tishk International University.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

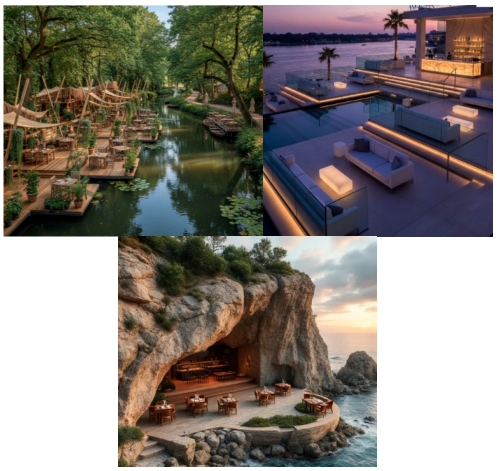
Acknowledgments: The researcher extends sincere gratitude to all third-stage students of the Architectural Engineering Department at Tishk International University during the academic year (2025–2026) for their commitment and cooperation. Appreciation is also expressed to the teaching staff for their valuable assistance throughout the investigation and data collection process, as well as to the independent panel of judges for their important contributions. Also, the author would like to clarify that during the preparation of this manuscript, he used Microsoft Copilot 365 for language refinement, polishing, and grammar checking. Following the use of this tool, the author reviewed and edited the output as necessary. The author takes full responsibility for the integrity and final content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

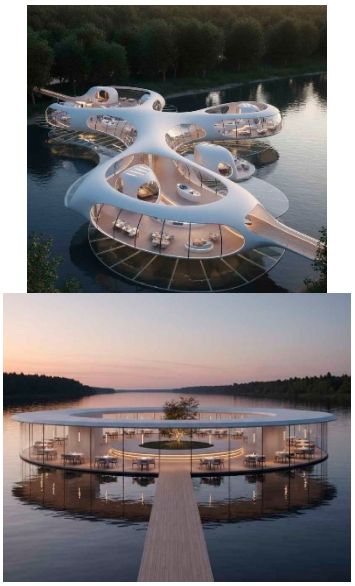
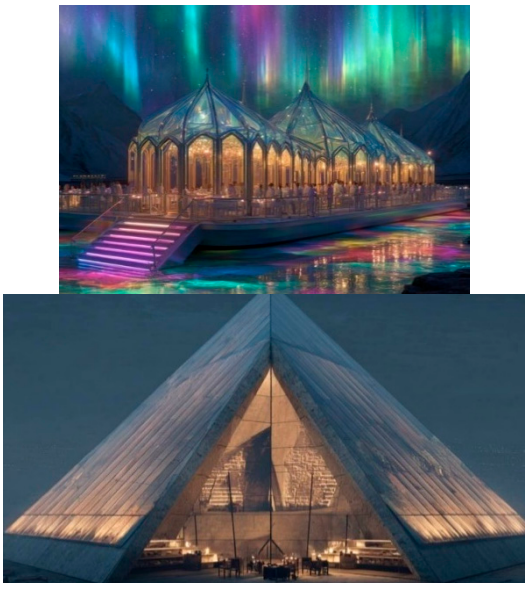
Appendix A

The images shown represent the AI-generated outputs produced from each student's submitted prompt(s). R tiers are indicated parenthetically for each student: S9 and S7 produced Sparse-tier prompts ($R < 30$), while students (S1, S3, S4, S11, S12) produced Moderate-tier prompts ($30 \leq R \leq 59$). Students (S10, S15) produced Rich-tier prompts ($R > 59$). The progression from Sparse to Moderate to Rich is visually evident in the increasing architectural coherence, material specificity, and compositional clarity of the outputs.

No.	Prompt	Result
S1	P1: A fluid, organic restaurant inspired by the gentle flow of water; its curving forms mimic waves shaping the shore, blending architecture and nature. P2: A riverside restaurant with a parasitic-inspired design, featuring a ribbed wooden extension that clings to the main stone structure. P3: A bold parasitic form anchored in stone; the design emerges from the landscape, blending raw nature with sharp geometry in harmony.	

No.	Prompt	Result
S3	P1: Floating nature restaurant, wooden decks, open air, surrounded by trees and a river. P2: River sky lounge stepped terrace overlooking river, modern and elegant form. P3: River life restaurant, inspired by fishing culture	
S7	P1: A design inspired by travel is a place for tourists, a place like a resort. P2: A warm, earthy design inspired by the cycle of nature, from soil to plate. P3: A modern, sophisticated restaurant set in the city, inspired by light, water, and glass.	
S9	P1: A restaurant like Babylon times. P2: A restaurant floated between timelines. P3: A restaurant inspired by ancient Silla myth.	

https://doi.org/10.3390/architecture6030155

No.	Prompt	Result
S10	P1: A bridge that connects two riverbanks invites you to dine above the water. Smooth, flowing forms that curve and twist to create cosy dining pods and open lounges, with metallic tones. P2: A minimalist circular restaurant floating on a calm river, connected by a slender wooden walkway. Smooth matte white curved roof, full-height glass walls, light natural wood interior.	
S15	P1: Floating crystal restaurant above a calm river, inspired by Aurora lights. Geometric design. P2: Pyramid-shaped restaurant blending ancient Egyptian architecture with subtle steampunk elements. Elegant futuristic fantasy design with golden lighting, metal details, and soft atmosphere.	

References

1. Taura, T.; Nagai, Y. *Concept Generation for Design Creativity: A Systematized Theory and Methodology*; Springer: London, UK, 2013. [CrossRef]
2. Liikkanen, L.A.; Perttula, M. Exploring problem decomposition in conceptual design among novice designers. *Des. Stud.* **2009**, *30*, 38–59. [CrossRef]
3. Wu, Y.; Weng, K.; Young, L. A Concept Transformation Learning Model for Architectural Design Learning Process. *Eurasia J. Math. Sci. Technol. Educ.* **2016**, *12*, 1189–1197. [CrossRef]
4. Doha, E. Concept generation in the architectural design process: A suggested hybrid model of vertical and lateral thinking approaches. *Think. Ski. Creat.* **2019**, *33*, 100589. [CrossRef]
5. Lawson, B. *What Designers Know*; Routledge: London, UK, 2004.
6. Al Maani, D.; Roberts, A. An Attempt to Understand the Design Studio as a Distinctive Pedagogical Setting. *Int. J. Des. Educ.* **2023**, *17*, 31–44. [CrossRef]

7. Cross, N. *Design Thinking*, 2nd ed.; Bloomsbury Academic: London, UK, 2013.
8. Babangida, D.; Khaidzir, K. Concept of Creativity and Innovation in Architectural Design Process. *Int. J. Innov. Manag. Technol.* **2015**, *6*, 16–20. [CrossRef]
9. Cross, N.; Dorst, K. Creativity in the design process: Co-evolution of problem-Solution. *Des. Stud.* **2001**, *22*, 425–437. [CrossRef]
10. Cinar, S.; Demiroz, M. Integrating Text-to-Image AI in Architectural Design Education: Analytical Perspectives from a Studio Experience. *J. Des. Stud.* **2024**, *6*, 247–258. [CrossRef]
11. Horvath, A.; Pouliou, P. AI for conceptual architecture: Reflections on designing with text-to-text, text-to-image, and image-to-image generators. *Front. Archit. Res.* **2024**, *13*, 593–612. [CrossRef]
12. Alamasi, R.; Asfour, O.S. Applications of Generative AI in Architectural Design Education: A Systematic Review and Future Insights. *Digital* **2026**, *6*, 6. [CrossRef]
13. Yiannoudes, S. Text-to-Image Generative Artificial Intelligence as Conceptual Mechanism for Architectural Design. *J. Artif. Intell. Archit.* **2025**, *4*, 70–88. [CrossRef]
14. Fang, C.; Zhu, Y.; Fang, L.; Long, Y.; Lin, H.; Cong, Y.; Jia Wang, S. Generative AI-enhanced human-AI collaborative conceptual design: A systematic literature review. *Des. Stud.* **2025**, *97*, 101300. [CrossRef]
15. Koh, I. AI Architecture Studio Without Architects. In *Architectural Intelligence in the Age of Artificial Intelligence*; Gerber, A., Franck, O., Mieskes, M., Eds.; Architekturen: Wetzlar, Germany, 2025; pp. 228–240. [CrossRef]
16. Rozdolska, A. Artificial intelligence as part of future practices in the architect's work: Midjourney generative tool as part of a process of creating an architectural form. *Architectus* **2022**, *3*, 95–104. [CrossRef]
17. Ceylan, S. Artificial Intelligence in Architecture: An Educational Perspective. In *Proceedings of the 13th International Conference on Computer Supported Education, Online, 23–25 April 2021*; SciTePress: Setúbal, Portugal, 2021; Volume 1, pp. 100–107. [CrossRef]
18. Enjellina, B.; Rossy, A. A Review of AI Image Generator: Influences, Challenges, and Future Prospects for Architectural Field. *J. Artif. Intell. Archit.* **2023**, *2*, 53–65. [CrossRef]
19. Buthayna, E. Concept evolution in architectural design: An octonary framework. *Front. Archit. Res.* **2018**, *7*, 180–196. [CrossRef]
20. Al-Qemaqchi, N.T. The Transformation of Architectural Design Concepts During the Early Design Phase. *Int. J. Eng. Pedagog.* **2022**, *12*, 85–99. [CrossRef]
21. Taura, T.; Yamamoto, E.; Fasiha, M.Y.N.; Goka, M.; Mukai, F.; Nagai, Y.; Nakashima, H. Constructive simulation of creative concept generation process in design: A research method for difficult-to-observe design-thinking processes. *J. Eng. Des.* **2011**, *23*, 297–321. [CrossRef]
22. Cikis, S.; Ipek Ek, F. Conceptualization by Visual and Verbal Representations: An Experience in an Architectural Design Studio. *Des. J.* **2010**, *13*, 329–354. [CrossRef]
23. Do, E.; Gross, M. Thinking with diagrams in architectural design. *Artif. Intell. Rev.* **2001**, *15*, 135–149. [CrossRef]
24. Dogan, F. Architectural Design Students' Explorations through Conceptual Diagrams. *Des. J.* **2013**, *16*, 103–124. [CrossRef]
25. Bilda, Z.; Gero, J.; Purcell, T. To sketch or not to sketch? That is the question. *Des. Stud.* **2006**, *27*, 587–613. [CrossRef]
26. Ruckpaul, A.; Kriltz, A.; Matthiesen, S. Using Eye Tracking to Understand the Engineering Designers' Behavior in Synthesis-Driven: Analyzing Processes—Experiences in Study Design. In *Analyzing Cognitive Processes During Design: Proceedings of the HBiD 2014*; Meboldt, M., Matthiesen, S., Badke-Schaub, P., Lohmeyer, Q., Eds.; ETH Zürich: Zürich, Switzerland, 2014; pp. 1–8.
27. Avidan, Y.; Goldschmidt, G. Talking Architecture: Language and Its Roles in the Architectural Design Process. In *ICoRD'13: Lecture Notes in Mechanical Engineering*; Chakrabarti, A., Prakash, R., Eds.; Springer: New Delhi, India, 2013. [CrossRef]
28. Al-Qemaqchi, N.; Saeed, U. The alteration of designer logical performance during architectural design tasks. *Al-Rafidain Eng. J.* **2009**, *17*, 52–66. [CrossRef]
29. Al-Qemaqchi, N.; Ismaeel, E. The Transformation of Knowledge in Urban Design Studio. In Proceedings of the 2nd International Conference on Research in Science, Engineering and Technology (ICRSET'2014), Dubai, United Arab Emirates, 21–22 March 2014; pp. 167–173. [CrossRef]
30. Pasin, B. Rethinking the Design Studio-Centered Architectural Education. A Case Study at Schools of Architecture in Turkey. *Des. J.* **2017**, *20*, S1270–S1284. [CrossRef]
31. Llach, D. Sculpting Spaces of Possibility: Brief History and Prospects of Artificial Intelligence in Design. In *The Routledge Companion to Artificial Intelligence in Architecture*; As, I., Basu, P., Eds.; Routledge: London, UK, 2021.
32. Veloso, P.; Krishnamurti, R. Mapping generative models for architectural design. In *The Routledge Companion to Artificial Intelligence in Architecture*; As, I., Basu, P., Eds.; Routledge: London, UK, 2021.
33. 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]
34. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 10674–10685. [CrossRef]

35. Saharia, C.; Chan, W.; Saxena, S.; Whang, J.; Denton, E.; Ghasemipour, S.; Ayan, B.; Mahdavi, S.; Lopes, R.; Salimans, T.; et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv* **2022**, arXiv:2205.11487. [CrossRef]
36. Paananen, V.; Oppenlaender, J.; Visuri, A. Using text-to-image generation for architectural design ideation. *Int. J. Archit. Comput.* **2023**, *22*, 458–474. [CrossRef]
37. Oppenlaender, J. The Creativity of Text-to-Image Generation. In *Proceedings of the 25th International Academic Mindtrek Conference, Tampere, Finland, 16–18 November 2022*; ACM: New York, NY, USA, 2022; pp. 192–202. [CrossRef]
38. Han, S.; Fussell, S. Understanding User Perceptions and the Role of AI Image Generators in Image Creation Workflows. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan, 26 April–1 May 2025*; ACM: New York, NY, USA, 2025; pp. 1–17. [CrossRef]
39. Koo, T.; Li, M. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
40. Bin, Z. Halo Effect. In *The ECPH Encyclopedia of Psychology*; Springer: Singapore, 2024; pp. 1–2. [CrossRef]

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.