ArchiJury: Exploring the Capabilities of Vision-Language Models to Generate Architectural Critique

Selen Çiçek[1], Mehmet Sadık Aksu [2], Emre Öztürk[3], Kaan Bingöl[4], Gizem Mersin[5], Mustafa Koç[6], Oben K. Akmaz[7], Lale Başarır[8]

ORCID NO: 0000-0003-2489-2536[1], 0009-0004-7024-3304[2], 0009-0009-9937-4099[3], 0000-0001-7175-3198[4], 0009-0000-2295-353X[5], 0000-0001-8131-8878[6], 0000-0001-8620-6429[8]

[1,6] Istanbul Technical University, Graduate School, Department of Informatics, Architectural Design Computing, Istanbul, Turkey
[2,8] İzmir University of Economics, Faculty of Fine Arts and Design, Architecture, İzmir, Turkey
[1,2,3,4,5,6,7,8] Unified Methods of Artificial Intelligence (UMAI) Lab

Artificial Intelligence (AI) offers a potent opportunity to rethink architectural critique, in cases such as architectural design competitions. The challenge lies in capturing the interpretive depth required for design evaluation—an inherently human process that connects intuition, reasoning, and contextual sensitivity. Building on this premise, ArchiJury uses a domain-specific dataset, curated and validated by authors as domain experts, architects, to train a context-aware Visual-Language Model (VLM) capable of delivering a nuanced critique. The model development follows two distinct phases: an initial version (v1) explores feasibility through classification of visual architectural attributes, while the second phase (v2) evolves into a structure generating detailed critique texts guided by predefined criteria such as context, form, and programmatic considerations. The proposed model aims to bridge the gap between computational precision and the complexities of architectural judgment, offering a structured yet adaptable framework for utilizing AI in the evaluative aspects of design. Although still in its early stages, this work opens a pathway to complement traditional review processes with reliable, scalable, and context-sensitive feedback, laying a foundation for incorporating the patterns of tacit knowledge in architectural design into the review process.

Keywords: Architectural Critique, Artificial Intelligence (AI), Vision-Language Models (VLM), AI and Architectural Design, Architecture Competitions

165

JCoDe | Vol 6 No 1 | March 2025 | Ecological Intelligence in Computational Design| Çiçek, S., Aksu, M S., Öztürk, E., Bingöl, K., Mersin, G., Koç, M., Akmaz, O. K., Başarır, L.

Mimar Jüri: Görme-Dil Modelleri ile Mimari Tashih Üzerine bir İnceleme

Selen Çiçek[1], Mehmet Sadık Aksu [2], Emre Öztürk[3], Kaan Bingöl[4], Gizem Mersin[5], Mustafa Koç[6], Oben K. Akmaz[7], Lale Başarır[8]

ORCID NO: 0000-0003-2489-2536[1], 0009-0004-7024-3304[2], 0009-0009-9937-4099[3], 0000-0001-7175-3198[4], 0009-0000-2295-353X[5], 0000-0001-8131-8878[6], 0000-0001-8620-6429[8]

[1,6] Istanbul Technical University, Graduate School, Department of Informatics, Architectural Design Computing, Istanbul, Turkey
[2,8] İzmir University of Economics, Faculty of Fine Arts and Design, Architecture, İzmir, Turkey
[1,2,3,4,5,6,7,8] Unified Methods of Artificial Intelligence (UMAI) Lab

Günümüz tasarım pratiğini radikal şekilde dönüştürmeye başlayan üretken Yapay Zeka (YZ) modelleri, tasarım sürecinin derinlemesine değerlendirilmesi ve geliştirilmesi için kritik bir öneme sahip olan mimari eleştiri için önemli bir potansiyel sunmaktadır. Özellikle, mimari tasarım yarışmaları gibi yoğun katılımcı sayısına sahip, kapsamlı ve tutarlı mimari eleştirilerin elzem olduğu çerçevelerde mimari kritiğe ulaşmak büyük bir zorluk oluşturmaktadır. Bu noktada çalışma, Görme Dil Modelleri olarak bilinen bir yapay zeka modeli mimarisini, tasarım problemlerini sorgulayarak, üretilen mimari çözümlere yorum ve mimari eleştiri geliştirmek üzere kullanılmasını önceleyen bir çerçeve önermektedir. Mimari tasarım pratiklerinde YZ araçları daha çok üretim, görsel temsil ve optimizasyon gibi somut çıktılar elde etmek için kullanılsa da, mimari eleştiri gibi sezgisellik, sorgulama ve bağlamsallık gerektiren alanlarda henüz sınırlı bir kullanım alanına sahiptir. Araştırma kapsamında önerilen YZ modelinin mimari eleştirinin sezgisel ve yoruma dayalı, nicel veriler ile ölçülemeyen boyutlarına entegre edilerek, tutarlı ve ölçeklenebilir eleştirilerin geliştirilmesi amaçlamaktadır. Önerilen model, hem bağlam duyarlılığı hem de mimari değerlere uygunluğu sağlamak adına yazarlar, alan uzmanları, tarafından tasarlanmış bir veri seti ile eğitilmiştir. Modelin geliştirilmesi safhası, iki temel aşamadan oluşmaktadır. İlk aşama olan "v1," görsel mimari özelliklerin (örneğin, geleneksel veya çağdaş, açısal veya organik formlar gibi) ikili sınıflandırmasını inceleyerek, çalışmanın ikinci aşamasında geliştirilen model mimarisinin tanımlanan araştırma problemi karşısında uygulanabilirliğini test etmeyi amaçlamaktadır. İkinci aşama olan "v2"de ise model mimarisi, önceden tanımlanmış değerlendirme kriterlerini (bağlam, ölçek, tasarım stratejileri, programatik ilişkiler vb.) kullanarak kapsamlı ve detaylı metinsel eleştiriler üretmek üzere geliştirilmiştir. İlk aşamada elde edilen sonuçların değerlendirilmesinin ardından; ikinci versiyonda model, genişletilmiş bir görsel veri seti ve uzman değerlendirmesiyle elde edilen mimari yorumlar ile eğitilerek, modelin kapsamlı ve tutarlı eleştiriler üretme kapasitesi artırılmıştır. Bu süreçte, modelin ürettiği her eleştiri, doğruluk ve tutarlılık açısından alan uzmanları tarafından gözden geçirilmiş ve revize edilmiştir. Çalışma kapsamında elde edilen sonuçlar, Görme Dil Modellerinin geleneksel jüri süreçlerini yapılandırılmış, ölçeklenebilir ve bağlam duyarlı eleştirilerle destekleyerek mimari tasarım pratiği ve yapay zeka arasındaki diyaloğu geliştirme potansiyeline sahip olduğunun altını çizmektedir.

Anahtar Kelimeler: Mimari Eleştiri, Yapay Zeka, Görme Dil Modelleri (GDM), Yapay Zeka ve Mimari Tasarım, Mimari Tasarım Yarışmaları

1. INTRODUCTION

Architectural critique is the cornerstone of the architectural design practice, opening the discussion space for expanding multi-dimensional aspects of the given design problems, to guide and refine the continuous process of design (Lymer, 2009; Fischer et al., 1993). However, obtaining architectural critique during off-studio hours is often limited by availability and accessibility, making it challenging to integrate consistent feedback into evolving processes of design (Luther et al., 2015). Nowhere is this more evident than in the context of architectural design competitions, which often represent the pinnacle of professional practice. These competitions require rigorous evaluation of entries that are innovative yet feasible and contextual (Rönn, 2011). Given that submissions are often numerous-sometimes in the hundreds, and typically occur within time-constrained competitions, it can be challenging for jurors to provide fair and comprehensive critiques, making it difficult to ensure that strong designs receive the recognition they deserve (Frederickson, 1990).

As artificial intelligence (AI) continues to transform architectural design, its potential to assist in competition review processes presents a compelling area of exploration. While in the current state of the art AI has been employed dominantly in architecture practice for generative design, visualization, and optimization tasks (Salem et al., 2024; Li et al., 2024), its application in the interpretive domain of critique remains underexplored. Because the acknowledgement of AI as a "black box" (Adadi & Berrada, 2018) poses significant challenges when applied to architectural critique since the problem goes beyond simply optimizing or generating output within well-defined parameters and metrics. Rather than the tangible aspects of a design – such as its spatial configuration or structural integrity — architectural critique also requires consideration of the intangible aspects, including cultural context, experiential resonance, and design intent (Güzer, 1994). Since it is a deeply interpretive process, intertwining intuition, reasoning, and contextual sensitivity—qualities that traditionally stem from human expertise, developing an AI model that can provide nuanced critique, these limitations must be addressed by combining computational precision with transparency and interpretive depth.

In light of this discussion, this research introduces an AI-driven architectural critique trained using Vision-Language Models (VLMs) to review architectural designs comprehensively. While the proposed ArchiJury research marks a rigorous initial step, it is only a fragment of the broader challenge of embedding AI in interpretive, intuitive and evaluative aspects of design practice. The main aim of this research presented here is to develop a synthetic architectural review model that can help assess entries in architectural design competitions. We hypothesize that the proposed VLM model trained with domain expert validated architectural image and review datasets, has potential to enhance traditional jury methods with structured, consistent, and scalable critiques of designs based on architectural principles. To this end, the study seeks to explore whether a domain-specific VLM can accurately generate structured critiques for architectural images, particularly by addressing contextual, formal, and programmatic criteria. This guiding question informs the methodological approach and evaluation framework detailed in the subsequent sections.

To draw a theoretical framework for the discussion in Section 1.1., the recent research in the computational design literature that delve into the role of AI in architectural design as evaluator are revisited. In section2, the computational architecture of the Vision Language Models (VLMs) are explained and the related literature that utilizes VLM models in the scope of architectural design are revisited. Starting from the overview of the methodology in Section 3, the development phases of the model are explained regarding two complementary versions. The utiliziation technique of the VLM architecture in the scope of the research is explained, which enabled us to generate architectural reviews based on architectural images. In Section 3.1 the training process of the initial version of the model is discussed, that focuses on classification tasks to identify key attributes of architectural designs from the provided architectural building photographs, that lays the groundwork for more complex interpretive capabilities. Subsequent iterations during the development phase of the second version of the model are discussed in Section 3.2, in terms of introducing comprehensive critique generation, enabling the model to deliver detailed feedback on contextual relationships, formal qualities, and programmatic considerations. In Section 4, the outputs of both model versions are compared and discussed in terms of the quality of the generated architectural critiques, revealing the limitations of the current model that outlines the further research investigation paths.

1.1. AI models for architectural design evaluation

The integration of Artificial Intelligence (AI) in architectural design evaluation is rapidly gaining traction. While much of the existing research focuses on AI as a design generator, there is a growing interest in exploring its potential as a critical evaluation tool. This section briefly reviews and highlights several key studies that examine computational tools and AI's role in assessing architectural designs, emphasizing their methodologies, limitations, and implications.

One notable study by Guzelci and Sener (2019) introduced an entropy-based model to evaluate projects submitted to architectural design competitions. This model considers various factors such as aesthetics, functionality, context, and innovation, assigning weights to each criterion to create a more objective assessment framework. Similarly, Luther et al. (2015) developed the CrowdCrit platform, a user-driven feedback system designed to gather diverse perspectives during the design process. This approach allows designers to share their projects with a broader audience and receive constructive feedback. Despite its value in democratizing critique, the platform lacks advanced contextual analysis capabilities, restricting its scope to user comments and ratings. Another important contribution comes from Wu et al. (2020), who developed an AI model to classify the visual characteristics of building facades. While this study demonstrated the model's efficacy in evaluating aesthetic properties, it failed to account for broader contextual and user-centric considerations. Its narrow focus on measurable visual elements underscores the need for more comprehensive evaluation approaches. Sanalan (2022) explored the transformative role of AI and big data technologies in architectural design processes. This study highlighted how these technologies are reshaping design workflows, enabling faster decision-making and more efficient collaboration. However, the research also emphasized the necessity of integrating qualitative and human-centric insights into AI-driven evaluations to avoid overly mechanistic assessments.

Despite these advancements, a common limitation of these studies is their dependence on quantifiable parameters. Architectural critique often involves complex, subjective judgments that extend beyond data-driven metrics. For instance, evaluating how a design aligns with its environmental context or enhances user experience requires a combination of intuitive and empirical approaches. To address these

ArchiJury: Exploring the Capabilities of Vision-Language Models to Generate Architectural Critique

gaps, this paper proposes a context-sensitive Visual-Language Model (VLM) as an alternative framework for architectural evaluation. Building on the strengths of multi-factor approaches like Guzelci and Sener's model (2018) , VLM incorporates the diversity and contextual richness highlighted by Luther et al. while maintaining technical rigor. Unlike existing methods, VLM prioritizes holistic assessments by integrating qualitative feedback with quantitative analysis, offering a more balanced evaluation of architectural projects. By treating AI not merely as a tool for generating designs but as a critical evaluator, VLM opens new possibilities for advancing architectural practice and fostering more nuanced critiques.

2. VISION LANGUAGE MODELS (VLMs)

Vision-Language Models (VLMs) represent a significant advancement in artificial intelligence, integrating visual and textual data to interpret, analyze, and generate context-aware outputs (Ghosh et al., 2024). VLMs function by linking computer vision techniques with natural language processing (NLP) algorithms, enabling systems to process visual inputs and produce textual interpretations (Bordes et al., 2024). Computational architecture of VLM models advanced rapidly, employing transformer architectures and self-attention mechanisms to learn joint representations of text and visual inputs, and are generally categorized into vision-language understanding models, multimodal input to unimodal output models, and multimodal input-output models (Ghosh et al., 2024). Despite these innovations, key challenges in VLM development persist, including data selection, architecture design, and training methods, alongside broader concerns regarding multimodal fusion, interpretability, reasoning, and ethical implications (Laurençon et al., 2024).

The architecture of the proposed VLM follows a modular design to facilitate efficient training and critique generation, incorporating a dual-stream encoder for processing visual and textual inputs, with the visual Vision Transformer (ViT) and the textual encoder employing a transcoder utilizing a pre-trained Transformer-based architecture (Marafioti et al., 2024). A cross-modal attention mechanism aligns visual and textual embeddings, forming the core of the model's interpretive capabilities. For critique generation, a decoder synthesizes structured

textual outputs addressing evaluative dimensions such as contextual relationships, formal characteristics, and programmatic considerations.

2.1. Utilization of VLMs in the Architectural Design Domain

The integration of Vision-Language Models (VLMs) into architectural design represents a novel and evolving research frontier. While generative AI models such as diffusion models and GANs have been widely employed for image generation and visualization tasks, the use of VLMs for interpretive and evaluative purposes in architecture remains relatively underexplored. However, recent scholarly efforts indicate an increasing interest in leveraging the multimodal reasoning capacities of VLMs within the design domain.

One pioneering study by Chen et al. (2024) introduced LLM4DESIGN, an automated multi-modal system that integrates VLMs with multi-agent systems and Retrieval-Augmented Generation (RAG) to facilitate architectural and environmental design processes. The research emphasizes generating coherent, multi-illustrated, and multi-textual design schemes that align with narrative storytelling and objective design presentations. By leveraging VLMs, the system is capable of analyzing design requirements alongside site conditions to produce context-aware architectural outputs, indicating the potential of VLMs for expanding the scope of design automation and critique.

Similarly, Jang and Lee (2023) explored the integration of large pre-trained language models with Building Information Modeling (BIM) systems, highlighting how VLM architectures can mediate the interaction between architectural data and design interpretation. Their approach utilized XML data formatting to translate between textual inputs and BIM models, revealing a pathway for VLMs to dynamically engage with structured architectural information and contribute to iterative design processes. Although the focus remains on system interoperability, the study underscores the capacity of VLMs to interpret and contextualize architectural information within computational frameworks.

Additionally, Galanos et al. (2023) introduced Architext, a language-driven design tool that employs large-scale language models to

generate architectural layouts based on natural language prompts. While primarily focused on conceptual generation rather than critique, the research illustrates the potential of VLMs to mediate between textual design intentions and spatial configurations. The study demonstrated the system's ability to produce valid residential layouts from minimal textual input, suggesting promising avenues for integrating VLMs into early-stage design ideation.

These emerging studies collectively illustrate the gradual incorporation of VLMs into architectural practice, particularly in contexts requiring the translation of complex, multimodal information into actionable design insights. However, the direct application of VLMs for nuanced architectural critique—particularly in assessing the contextual, formal, and programmatic dimensions of architectural proposals—remains largely underdeveloped. Addressing this gap, the present study advances the discourse by proposing a domain-specific VLM framework explicitly tailored for architectural critique. Unlike prior applications that primarily focus on generative design or descriptive tasks, this research emphasizes the interpretive and evaluative capacities of VLMs. The model is designed to generate structured critiques that engage with the contextual relationships, formal qualities, and programmatic considerations inherent in architectural proposals.

3. TRAINING THE MODELS FOR ARCHITECTURAL CRITIQUE: OVERVIEW

This section revisits the research's overarching aim of using Vision-Language Models (VLMs) as synthetic jury members in architectural critique. It provides an overview of the research workflow while addressing the rationale for training a domain-specific AI model.

The rationale for developing a domain-specific AI model lies in the inherent limitations of general-purpose language models when applied to architectural critique. General-purpose language models often suffer from hallucinations—producing plausible yet incorrect outputs—due to their training on broad, multi-domain datasets (Zhang et al., 2023). This issue, exacerbated by computational parameters such as temperature, underscores the need for a targeted approach (Mittal et

al., 2024). Furthermore, generic models have limited ability to understand contextual sensitivity and nuanced concepts such as spatial relationships (Gokhale et al., 2022), cultural relevance (Shen et al., 2024), and design intent for architectural critique. Besides, these models also contain biases from their datasets which are unreliable in special contexts (Mehrabi et al., 2021). To overcome these biases, this proposed methodology adopts a comprehensive strategy: training the model on a domain-specific, context-aware dataset validated by six architects with a minimum of five years of experience; embedding architectural standards and principles are embedded into the dataset annotations; and a targeted task framework is defined to guarantee precise, structured outputs. This integrated approach not only mitigates hallucinations and biases but also ensures consistent, scalable, and professionally relevant critiques.

The research methodology is presented in two main phases of the developmental process, each with its own objectives and the sets of data used for training the goals **(Figure 1)**. The first version, v1, takes an initial approach with classification tasks to sort out architectural images into different categories based on certain formal and stylistic characteristics. As the initial step validated the feasibility of the proposed AI-driven critique framework, the subsequent version, v2, aims to train a more advanced model that generates comprehensive textual critiques based on the pre-defined evaluation criteria with the specific task definition.

The VLM employed in this study was developed as a domain-specific model tailored for architectural critique, leveraging a curated dataset validated by architectural domain experts to ensure reliable and context-sensitive outputs. For this purpose, the SmolVLM framework developed by Hugginface was chosen due to its lightweight architecture, open-source licensing, high performance-to-size ratio, and low training cost. SmolVLM builds upon the Idefics3 architecture, introducing enhancements such as a compact SmolLM2 1.7B language model and a visual processing layer employing a pixel shuffle strategy that compresses visual information ninefold, significantly improving efficiency compared to prior models. This strategy optimizes visual inputs of 384x384 pixels with 14x14 internal patches, reducing memory usage while maintaining high performance (Huggingface, 2024). The

173

model's tailored design aligns with the resource constraints of this study, allowing it to be fine-tuned effectively using a single RTX 4090 GPU. Techniques like Quantized LoRA (QLoRA) further optimized training efficiency (Dettmers et al., 2023), making SmolVLM particularly suitable for this application.

The training workflow involved two iterative phases: Version 1 focused on classification tasks using a dataset of 1,589 architectural images, supplemented by synthetic images generated through text-to-image diffusion models, and Version 2 expanded this approach with a dataset of 12,320 textual critiques linked to architectural images, transitioning the model to generate detailed, multi-dimensional architectural reviews.
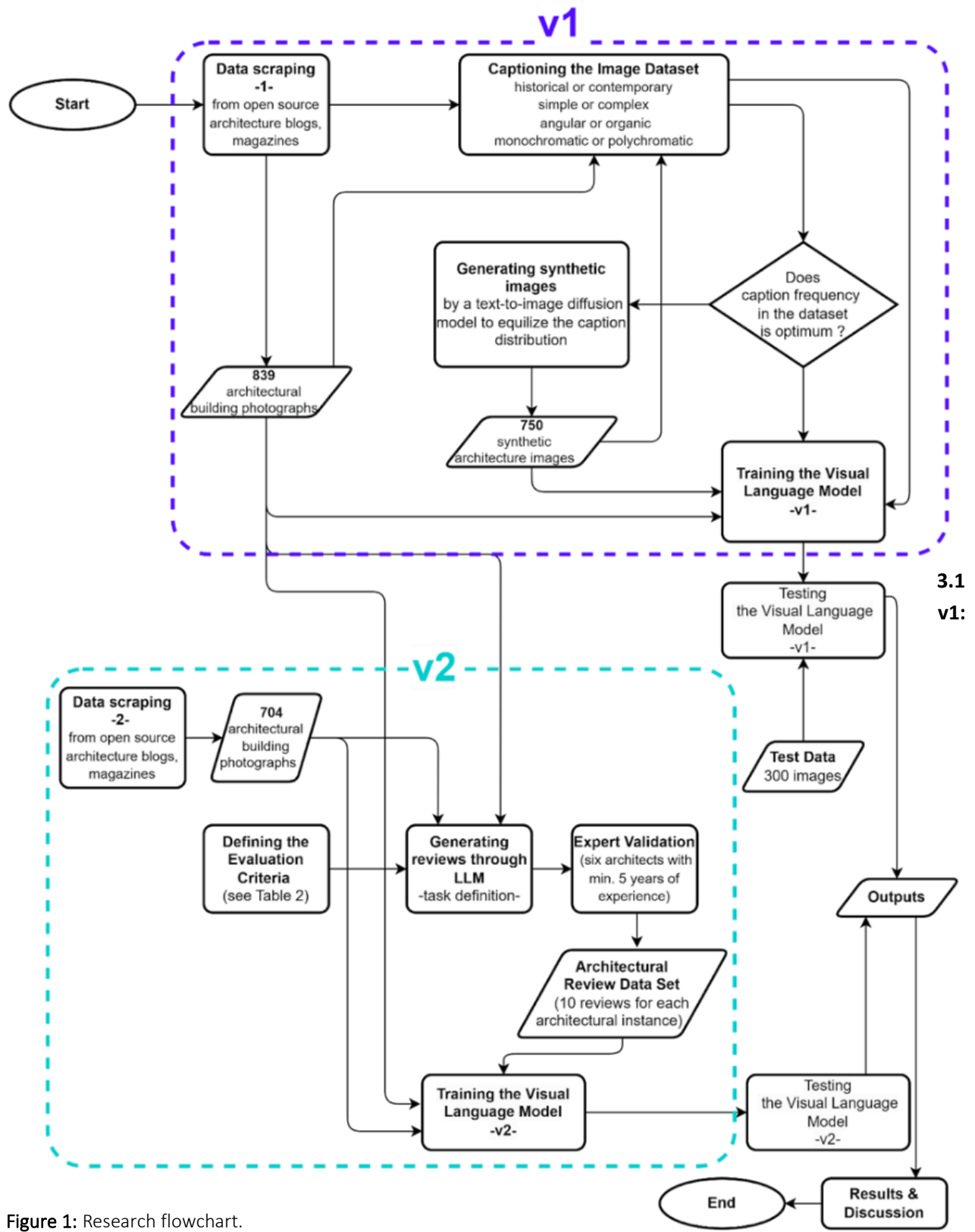
174

JCoDe | Vol 6 No 1 | March 2025 | Ecological Intelligence in Computational Design| Çiçek, S., Aksu, M S., Öztürk, E., Bingöl, K., Mersin, G., Koç, M., Akmaz, O K, Başarır, L.

Figure 1: Research flowchart.

ArchiJury: Exploring the Capabilities of Vision-Language Models to Generate Architectural Critique

Classification Based Descriptive Model

The initial version of the model, v1, was designed as a simple prototype to see the feasibility of the concept. Thus, a simple tagging system was utilized, concentrating on ease of implementation and evaluation. The goal was to determine whether the approach would be effective enough to warrant more sophisticated techniques being developed.

The v1 dataset consisted of 839 architectural images collected from a combination of international and local architectural websites. Well-known architecture blogs i.e. ArchDaily, Dezeen and Architectural Digest made up the international segment, while the Turkish platform Arkitera provided the local content. The data collection process was a combination of automated and manual efforts. Although a small portion of the dataset was curated manually, the vast majority of the images were collected through custom data scraping scripts, with a focus on Arkitera's massive project archives. For consistency, the first two images in each project's page were retrieved as the visual representatives of that specific architectural project. After the initial scraping, we removed images containing architectural drawings, presentation boards, noisy, deformed, ultra-wide perspective, and aerial perspective images to guarantee that the dataset contained only proper visuals of architectural structures. **Figure 2** displays the collected dataset instances partially.
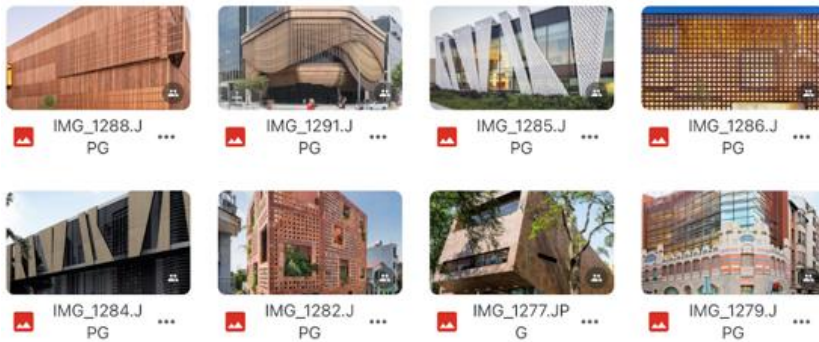
Figure 2: Examples from collected image dataset that contains architectural building photographs.

After the scraping and gathering processes, the images were subjected to a structured captioning process. Six domain experts, also architects, evaluated and captioned each image. The captioning was done based on visual characteristics of the images alone, without considering the context of the architectural design. **(Figure 3)**

The primary objective was to assess the baseline capabilities of the Vision-Language Model (VLM) in classifying architectural images based on simple, yet fundamental, visual characteristics. To ensure clarity and minimize interpretive complexity, we designed a controlled training approach using predefined caption categories that are visually discernible and fundamental to architectural analysis.

The selected classification criteria were based on four core visual attributes:
Form: Categorized as angular or organic, reflecting the geometric nature of the architectural form.
Style: Defined as contemporary or historic, indicating the temporal and stylistic context of the design.
Color Palette: Distinguished as monochromatic or polychromatic, referring to the visual character of the facade's color composition.
*Spatial Organization:* Classified as simple or complex, capturing the arrangement and complexity of spatial elements observable in the facade composition.

These categories were chosen for their clarity, objectivity, and relevance to architectural discourse. Each attribute represents a fundamental aspect of visual analysis that can be consistently identified from single images. For a single image of a building, the 16-class multi-class captions, formed by combinations of whether historical or contemporary, simple or complex, angular or organic, and monochromatic or polychromatic, were chosen for their ability to capture core architectural characteristics while being straightforward to differentiate.

Although the v1 dataset was technically suitable for training the model, there were problems with the distribution of captions in the data. As images were obtained from contemporary architectural blogs, the dataset was highly concentrated on contemporary buildings as expected. In addition to the single-caption imbalances, certain caption combinations also appeared disproportionately often. For instance, the most frequent combination was contemporary, complex, organic, monochromatic, while the least frequent combination was historical, complex, organic, polychromatic. **(Figure 4)**

ArchiJury: Exploring the Capabilities of Vision-Language Models to Generate Architectural Critique

A	B	C	D	E
NO	Contemporary-Historical	Simple-Complex	Angular-Organic	Monochromatic-Polychromatic
	Is the design contemporary or historical?	Does the structure appear simple or complex?	Does the design feature angular lines or organic curves?	Does the design use a monochromatic or polychromatic color palette?
	Comtemporary	Simple	Angular	Monochromatic
0	Contemporary	Complex	Angular	Monochromatic
1	Contemporary	Simple	Angular	Monochromatic
2	Contemporary	Complex	Angular	Monochromatic
3	Contemporary	Complex	Organic	Monochromatic
4	Contemporary	Simple	Angular	Monochromatic
5	Contemporary	Simple	Angular	Monochromatic
6	Contemporary	Complex	Organic	Monochromatic
7	Contemporary	Complex	Angular	Monochromatic
8	Contemporary	Complex	Angular	Monochromatic
10	Contemporary	Simple	Organic	Monochromatic
12	Contemporary	Simple	Angular	Polychromatic
13	Contemporary	Complex	Organic	Monochromatic
14	Contemporary	Complex	Angular	Monochromatic
15	Contemporary	Simple	Angular	Monochromatic

Figure 3: The spreadsheet used by the domain experts for the captioning the architectural instances in the dataset.
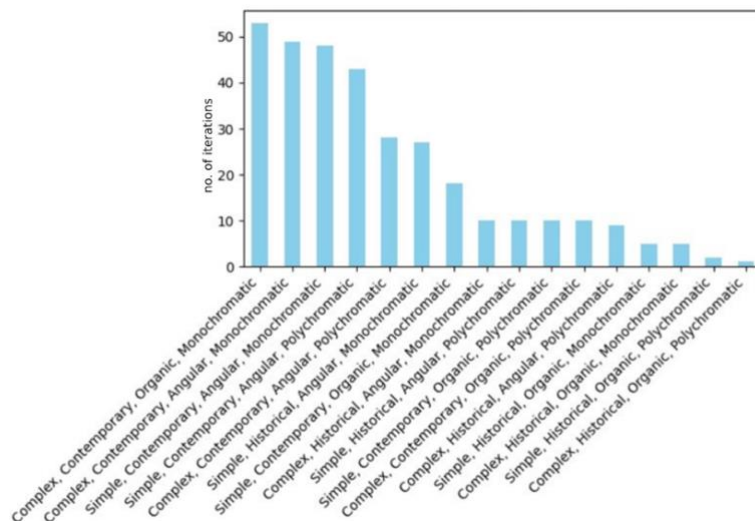
Figure 4: The distribution graph of the captions combinations in the initial version of the dataset.

This imbalance posed a risk of having a biased model, as certain architectural captions and combinations would dominate during training. To address this issue, we enhanced our dataset through a synthetic data generation method. Using text-to-image diffusion models, we created architectural building images with the less frequent caption combinations by incorporating these combinations into the prompts given to the text-to-image diffusion models. While the generated images were synthetic and do not exist in the real world, they increased the diversity of the dataset and provided the caption frequency and combinations balance.

178

In total, 750 synthetic images were generated, specifically from the underrepresented caption combinations. **Table 1** displays an exemplary selection of the synthetic architectural images and their text prompts generated by using accessible text-to-image diffusion models i.e. Midjourney, Prome AI. After generation, the synthetic images were also captioned correspondingly to maintain consistency. These synthetic images were then combined with the original scraped data to create a final v1 dataset of 1,589 images. The following comprehensive dataset provided a more balanced and representative foundation for training the first version of our model.

Table 1: Exemplary selection of the synthetic data included in the architectural images dataset.

Text-prompts	Generated Images
A simple, historical cultural pavilion with sharp, angular geometry and a vibrant polychromatic façade. The design features bold, triangular rooflines and color panels that shift from deep reds to soft blues, contrasting with the historical stonework. The angular shapes create a dynamic interplay of light and shadow, emphasizing the modern use of color within a historical context.	
Exterior architectural photography, Contemporary organic architecture with fluid, seamless forms, monochromatic material palette of raw concrete and warm-toned wood, natural stone accents, minimal yet bold design, soft transitions between surfaces, open and expansive spaces, calm and earthy aesthetic.	

During the training process of version 1, the pre-trained SmolVLM model was utilized. This decision was primarily driven by the high costs associated with training large-scale Vision-Language Models (VLMs) from scratch. Instead, a strategy was adopted to select pre-trained models suited to the target task and specialize them through fine-tuning with customized datasets. One of the primary challenges encountered during the training process stemmed from the dataset's classification-based structure, which relied on predefined categories. While the input images encompassed significant diversity and varied data types, the insufficient representation of labels (categories) to match this diversity posed difficulties for the model's learning process. The test results of the initial version of the model, v1 are displayed in **Figure 5** partially. The results of these experiments ultimately highlighted the need for a transition to version 2, accompanied by the preparation of a more comprehensive and balanced dataset, leading to concrete steps in this direction.
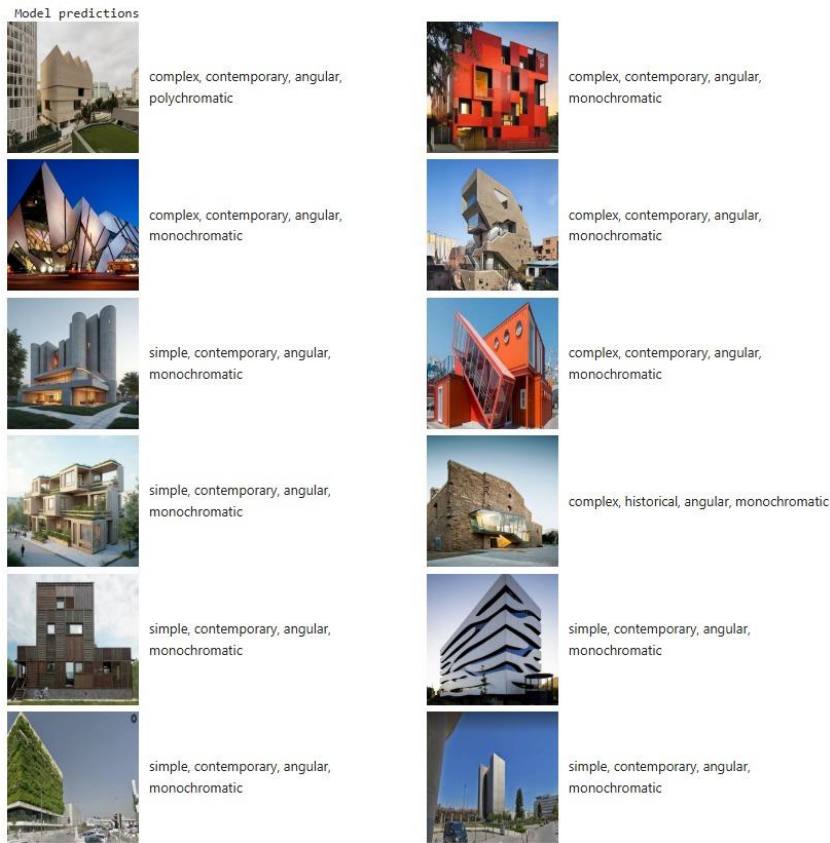
3.2. v2: VLM based Comprehensive Review Model

After confirming the feasibility of our approach with v1, to improve the performance of the model, computational architecture of the VLM model is altered from generating multi-class captions to comprehensive and descriptive architectural review structures. This improvement required both a larger dataset and a new approach for captioning. To achieve this, the initial image data set is extended by using data scraping and acquisition techniques. Approximately 15,000 images from over 2,400 architectural projects are scraped from Arkitera's archives using scripts. To ensure the quality and relevance of the images a multi-step filtering process is implemented. First, script-based methods to eliminate images containing technical drawings, presentation boards, and infographics were used. After, a manual screening was conducted to remove noisy or deformed images, ultra-wide perspectives, and aerial perspectives. After these rigorous preparation steps, 704 new high-quality images were added to the dataset. These were combined with real-world architectural images from the initial dataset used in v1. However, synthetic generated

images from v1 were excluded, as the need to balance captions was no longer relevant in this version. As a result, the second dataset contains a total of 1,232 architectural images.

Unlike v1, where multi-class captions were employed, v2 adopted a text-based captioning approach. Every image in the dataset was captioned with textual descriptions to address various architectural evaluation criteria. These criteria included context, form, architectural style, design principles, scale, program, and structural system etc. **(Table 2)**. The captions were created synthetically with the help of large language models (LLMs). For each image, the language model was prompted to review the architectural features based on these specific criteria, producing detailed review comments. To improve diversity and the ability of the model to learn, ten alternative captions are generated for each image instance.

Table 2: Defined evaluation criteria to generate architectural reviews.

Evaluation Criteria	
Context	Relationship with surroundings, Urban density, Urban texture
Form	Angular, Orthogonal
Design Principles	Rhythm, Balance, Contrast, Integrity, Hierarchy, Symmetry Datum, Harmony, Repetition
Architectural Style	Traditional, Modern, Post-modern, Parametric, Brutalist, Minimal, Eclectic, Classic, Neo-classic, Futuristic, Deconstructivist, Gothic, Baroque, Art Nouveau ,Renaissance
Height	Low-rise (1-4 floors), Mid-rise (5-12 floors), High-rise (13+ floors)
Program	Residential, Commercial, Public, Mixed-use, Industrial, Institutional , Religious
Construction Strategy	Additive, Masonry, Frame
Architectural Scale	Pavilion, Building, Urban
Construction System	Reinforced Concrete, Steel Frame, Wooden Frame

To generate detailed textual captions for the dataset, a specialized task definition is tailored to have comprehensive architectural reviews from the large language model: "As an expert architect, provide a one-paragraph critique on the depicted architectural design, considering its relationship with the context, form, design strategy, style, scale, construction strategy, structural system, and sustainable approach." The task definition in the text prompt structure was important to make sure LLM to mention all the given architectural evaluation criteria, without enabling the model to hallucinate.

After generating synthetic architectural review comments utilizing the large language model, all the 12,320 reviews went through manual examination by a team of six domain experts, all of whom were architects with a minimum five years of experience. This step was important to validate the quality and reliability of the dataset

annotations. Poorly constructed, insufficient, or unclear comments were manually revised and improved. This meticulous process guaranteed that the textual annotations maintained a high standard of accuracy and relevance, aligning with the expectations of professional architectural review. The whole process created an annotated dataset that v2 was trained on, enabling it to generate comprehensive architectural review comments. **Table 3** illustrates the dataset that is used to train the VLM model over one of the examples. The reviews displayed in the table are initially generated by an LLM and revised and validated by the experts of the study in terms of the review structure regarding the given task definition and evaluation criteria.

Table 3: Example of validated review alternatives generated through LLM from the input architectural image dataset.

Provided Image	Review 1	Review 2
	The building makes a bold visual statement in the natural coastal landscape, with its vertical, monolithic form standing in stark contrast to the rugged rocks around it. Its form is streamlined, with smooth, layered panels that create a sense of rhythm and flow, despite its height. Design principles of contrast and emphasis are evident, as the white structure rises dramatically from the rock. The futuristic style complements its unique setting, offering an iconic addition to the coastline. The large-scale, vertical design appears to serve a residential or hospitality purpose. Likely utilizing a steel or concrete core for stability, the building is designed to withstand its exposed cliffside environment.	Rising sharply from the cliffside, this building's smooth, vertical layers create a striking dialogue with the rocky landscape. Its form is bold and sleek, with a flowing façade that enhances the sense of height and elegance. Design principles of emphasis and contrast are evident, with the bright white color and modern aesthetic standing out against the natural terrain. Its futuristic style gives it a unique character, aligning with themes of innovation. Towering in scale, it seems suited for high-end residential or hotel use. Reinforced concrete likely forms the core structure, while smooth cladding materials create the layered, monolithic effect, making it both visually striking and durable.

Training the model by implementing SmolVLM's fine-tuning process, enables AI to preserve the original architecture while adapting it to the specific task using a specialized dataset. This approach was cost-effective and leveraged pre-existing capabilities of the model in visual understanding and interpretation. A consistent task prompt—"Evaluate the building architecturally in terms of its context, form, design principles, style, scale, architectural size, architectural program, construction strategy, structural system, sustainability, and green architecture"—ensured focus on architectural critique. Fine-tuning was performed over 375 steps with a loss plateau observed at step 250, indicating optimal training convergence. **(Figure 6)**

Using QLoRA, the fine-tuning process was efficient, requiring only a single GPU with 24 GB VRAM, which reduced computational costs and allowed broader accessibility. Loss values, while indicative of the

model's learning process, are not the sole measure of performance for large generative models like VLMs; practical evaluation often involves direct interaction with the model. In this study, we tested the model using architectural images it had not encountered before, demonstrating its ability to generate coherent and nuanced critiques. The checkpoint at step 250 was selected as the optimal stopping point based on loss stabilization and evaluation results on diverse architectural examples, yielding positive outcomes.

Figure 6: Plot of the loss values during training of the VLM model.

Smoothed	Value	Step	Relative
0,0008	0,0001	375	6.057 hr
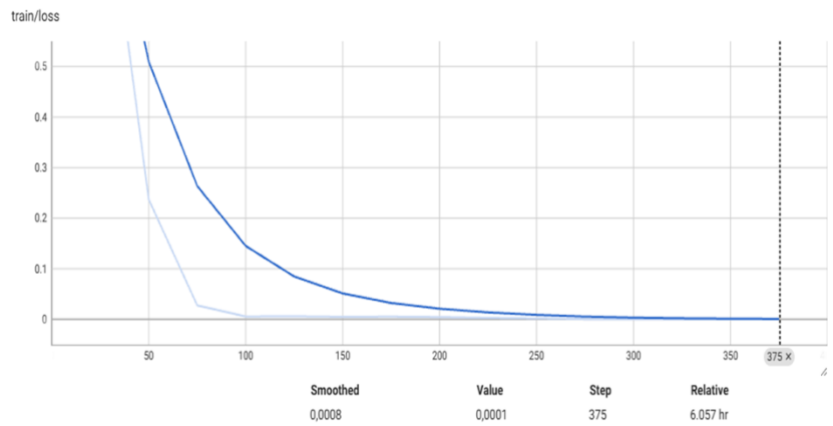
Table 4: Test results of the second version of the model, v2.

Table 4 displays two instance from the test results in terms of the architectural building photographs as input and the generated architectural review as output.

	Image (Input)	Review (Output)
(a)		This striking public building combines modern transparency with an organic design ethos, evident in its undulating roofline that mimics natural forms while framing the expansive glass façade below. The structure's scale and open interior suggest a civic or institutional function, likely a library or cultural center, where light and openness foster community interaction and knowledge sharing. The rhythmic arrangement of slender columns enhances verticality while maintaining an elegant balance and unity with the horizontal roof profile. The use of glass maximizes natural light and visual connection between the interior and exterior, emphasizing transparency and accessibility. The landscaped surroundings integrate seamlessly with the architecture, softening the formal grid of the plaza and creating a welcoming approach. The design's sustainable intent is evident in the thoughtful interaction with natural light and potential energy efficiency measures. Overall, this project masterfully blends form, function, and environmental sensitivity into a contemporary architectural landmark.

ArchiJury: Exploring the Capabilities of Vision-Language Models to Generate Architectural Critique

Table 4 (continued): Test results of
the second version of the model, v2.

| (b) | | This building's form and context respond effectively to its waterside location, where the sweeping wooden roof adds a natural, organic contrast to the reflective glass facade. The curvilinear roof creates rhythm and movement, while the orthogonal black facade elements offer balance and contrast. Modern in style, the building embodies contemporary architectural principles with a bold, sculptural form that commands the site without overpowering it. Its low-rise scale aligns with the surrounding waterfront structures, promoting cohesion. Likely designed as a public or mixed-use space, it fosters engagement and interaction. The construction seems to involve a steel frame supporting the expansive wooden canopy, illustrating advanced construction techniques. |

5. DISCUSSION AND INTERPRETATION OF GENERATED CRITIQUES

The initial v1 model demonstrated success in accurately identifying and categorizing architectural images based on predefined attributes, as reflected in the examples provided. By classifying images into clear categories such as "complex," "contemporary," "angular," or "monochromatic," the model established a reliable framework for recognizing architectural characteristics. However, this classification approach, while useful, simplifies architectural critique into distinct categories, limiting its capacity to engage with the interpretive and subjective aspects of design evaluation. While this approach offered a foundational structure for the critique process, it inadvertently stripped away the nuanced layers that define architectural judgment. Architectural critique extends beyond dichotomous classifications; it is an iterative and interpretive process that intertwines subjective and objective considerations.

By reducing architectural critique to classification labels, we risk overlooking the interconnected relationships between form, context, and experiential resonance. This simplification fails to account for the dynamic potentials and emergent qualities within a design that often defy clear categorization. The v2 model addresses this limitation by shifting from a multi-class framework to a more holistic and descriptive critique, capturing the complexities of architectural design within a context-sensitive structure.

184

JCoDe | Vol 6 No 1 | March 2025 | Ecological Intelligence in Computational Design| Çiçek, S., Aksu, M S., Öztürk, E., Bingöl, K., Mersin, G., Koç, M., Akmaz, O K, Başarır, L.

As displayed by **Table 4,** the v2 model test results demonstrate that generated reviews effectively highlight the contextual relationships between the architectural design and its environment. For instance, in the example **"b"** of the waterside structure, the critique captures the dialogue between the sweeping wooden roof and the reflective glass façade. This insight demonstrates the model's capacity to recognize and articulate the interplay between form and site, a critical aspect of architectural evaluation. The integration with the natural environment and the mention of cohesion with surrounding waterfront structures further reflect the nuanced understanding embedded in the model.

Both reviews in the table show the ability to evaluate design principles like rhythm, balance, and contrast. For example in instance **"a",** the undulating roofline in the public building example is noted for mimicking natural forms, creating a visual rhythm. This analysis aligns with key architectural critique practices where dynamic forms are evaluated for their aesthetic and functional implications. Similarly in instance **"b",** the contrast between the bright white structure and the rugged natural terrain in another example exemplifies the model's capacity to assess formal relationships within the architectural composition.

Both generated reviews in **Table 4** also briefly touch on sustainability and structural systems, which are critical components of modern architectural evaluation. In the example **"b"**, the use of a steel frame supporting the wooden canopy is identified, emphasizing the construction strategy and its alignment with advanced building techniques. Similarly in example **"a"** the maximization of natural light and the potential for energy-efficient measures in the public building further underline the model's capacity to evaluate sustainability features.

5.1. Limitations

While the proposed model marks an important step toward automating architectural critique, certain limitations in its current implementation warrant further discussion. At its core, the model functions as a descriptive tool, adept at analysing and articulating specific attributes of architectural images. This descriptive capability is valuable in identifying design qualities; however, it does not fully meet the requirements of an effective architectural critique. Critique in architectural practice is not confined to describing existing features but

ArchiJury: Exploring the Capabilities of Vision-Language Models to Generate Architectural Critique

also involves uncovering shortcomings and offering constructive suggestions for improvement.

Another significant limitation lies in the model's dependence on single-image inputs to generate critiques. Architectural evaluation, by its nature, is a holistic process that encompasses multiple dimensions of design. Reducing a complex architectural work to a single image oversimplifies its intricacies and omits vital information necessary for comprehensive critique. Evaluations based solely on façade images neglect critical aspects such as spatial organization, structural systems, material use, and functional programming. Plans, sections, elevations, and site analysis provide essential layers of information that reveal how a design functions and interacts with its context. Without these additional perspectives, the current model cannot fully capture the multidimensional nature of architecture, which significantly limits the depth and reliability of its critiques.

One critical aspect that remains unaddressed is the absence of scoring mechanisms in the current model. Human juries often evaluate architectural designs not only through qualitative critique but also by assigning quantitative scores to specific criteria, such as context, functionality, aesthetics, and innovation. These scores serve as an objective framework for comparing and ranking submissions, especially in design competitions. Without a scoring system, the model lacks the capability to quantify its assessments or offer a structured basis for comparison. Integrating a scoring mechanism would align the model more closely with real-world jury practices and enhance its utility as a decision-support tool in competitive settings.

5.2. Future Research Directions

As a future research plan of efforts to integrate AI-driven insights into architectural evaluation, we plan to develop a model that focuses on providing quantitative measures of an architectural image. The proposed model would utilize the existing framework of deep learning architecture and estimates and provides scores for four parameters: style, context, design principles, and total.
The proposed aesthetic model may utilize a ResNet-18 backbone which has been pre-trained in large datasets for effective feature extraction.

The architecture's last few layers are augmented with fully-connected layers fixed to output scores for different dimensions of scores. The model is trained end-to-end using a dataset annotated with ground-truth scores for each attribute. The predictions against the ground-truth are optimized by a mean squared error (MSE) loss function for increasing scoring accuracy.

To enhance explainability, the model is also required to include a Grad-CAM based explanation methodology that generates heatmaps for each predicted attribute. These heatmaps can indicate the parts of the image that influence the scoring decision the most and thus give various components of the model an understanding of how the model scores different aesthetic components. This explainability would offer us the opportunity to combine computational scoring with a more nuanced opinion of a human-like critique and therefore makes the model suitable for real-life and academic applications.

Acknowledgements

Author Contributions

All authors contributed equally to this article.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/access.2018.2870052

As, I., Pal, S., & Basu, P. (2018). Artificial intelligence in architecture: Generating conceptual design via deep learning. International Journal of Architectural Computing, 16(4), 306–327. https://doi.org/10.1177/1478077118800982

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H., . . . Chandra, V. (2024). An introduction to Vision-Language

modeling. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2405.17247

Denzler, J., Rodner, E., & Simon, M. (2016). Convolutional neural networks as a computational model for the underlying processes of aesthetics perception. In Lecture notes in computer science (pp. 871–887). https://doi.org/10.1007/978-3-319-46604-0_60

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient Finetuning of Quantized LLMS. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.14314

Fischer, G., Nakakoji, K., Ostwald, J., Stahl, G., & Sumner, T. (1993). Embedding critics in design environments. The Knowledge Engineering Review, 8(4), 285–307. https://doi.org/10.1017/s026988890000031x

Frederickson, M. P. (1990). Design Juries: A study in Lines of communication. Journal of Architectural Education, 43(2), 22–27. https://doi.org/10.1080/10464883.1990.10758556

Ghosh, A., Acharya, A., Saha, S., Jain, V., & CHadha, A. (2024). Exploring the Frontier of Vision-Language Models: A survey of current methodologies and future directions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2404.07214

Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., & Yang, Y. (2022). Benchmarking spatial relationships in Text-to-Image generation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2212.10015

Güzelci, O. Z., & Şener, S. M. (2019). An entropy-Based Design Evaluation Model for architectural competitions through multiple factors. Entropy (Basel, Switzerland), 21(11), 1064. doi:10.3390/e21111064

Güzer, C. A. G. (1994). The Limits of architecturalcritism: Architecture as a procress of represantation, commodification and legitimation [PhD Dissertation, Middle East Technical University]. https://open.metu.edu.tr/handle/11511/858

Laurençon, H., Marafioti, A., Sanh, V., & Tronchon, L. (2024). Building and better understanding vision-language models: insights and future directions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2408.12637

Li, C., Zhang, T., Du, X., Zhang, Y., & Xie, H. (2024). Generative AI models for different steps in architectural design: A literature review. Frontiers of Architectural Research. https://doi.org/10.1016/j.foar.2024.10.001

Luther, K., Tolentino, J., Wu, W., Pavel, A., Bailey, B. P., Agrawala, M., Hartmann, B., & Dow, S. P. (2015). Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. https://doi.org/10.1145/2675133.2675283

Luther, K., Williams, A., Hicks, J., & Dow, S. P. (2015). CrowdCrit: A crowdsourced approach to critique for design education. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 163–172.

188

JCoDe | Vol 6 No 1 | March 2025 | Ecological Intelligence in Computational Design| Çiçek, S., Aksu, M S., Öztürk, E., Bingöl, K., Mersin, G., Koç, M., Akmaz, O K, Başarır, L.

Lymer, G. (2009). Demonstrating Professional vision: The work of critique in Architectural education. Mind Culture and Activity, 16(2), 145–171. https://doi.org/10.1080/10749030802590580

Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., … Wolf, T. (2025). SmolVLM: Redefining small and efficient multimodal models.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Mittal, A., Murthy, R., Kumar, V., & Bhat, R. (2024). Towards understanding and mitigating the hallucinations in NLP and Speech. CODS-COMAD '24: Proceedings of the 7th Joint International Conference on Data Science & Management of Data, 489–492. https://doi.org/10.1145/3632410.3633297

Rönn, M. (2011). Architectural quality in competitions. A dialogue based assessment of design proposals. FormAkademisk - Forskningstidsskrift for Design Og Designdidaktikk, 4(1). https://doi.org/10.7577/formakademisk.130

Salem, A., Mansour, Y., & Eldaly, H. (2024). Generative vs. Non-Generative AI: Analyzing the Effects of AI on the Architectural Design Process. Engineering Research Journal (Shoubra), 53(2), 119-128. doi: 10.21608/erjsh.2024.255372.1256

Sanalan, A. (2022). The role of artificial intelligence and big data technologies in architectural design processes. Maltepe University Graduate Institute. Retrieved from Maltepe Open Access

Shen, S., Logeswaran, L., Lee, M., Lee, H., Poria, S., & Mihalcea, R. (2024). Understanding the capabilities and limitations of large language models for cultural commonsense. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2405.04655

SmolVLM - small yet mighty Vision Language Model. (2024). https://huggingface.co/blog/smolvlm

Wu, T., Liu, C., & Li, Y. (2020). Visual classification in architectural design: A neural network approach. Architectural Computing Journal, 18(2), 102–118.

Zhang, M., Press, O., Merrill, W., Liu, A., & Smith, N. A. (2023). How language model hallucinations can snowball. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.13534