buildings

MDPI

Article

The Use of Artificial Neural Networks in Architecture: Generating Depth-Aware and Explainable Building Facades from Floor Plans

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

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

Abstract

This study aims to position the use of artificial intelligence in architectural design, not merely as a tool for visual generation, but as an exploratory decision-support framework in which selected design relationships can be represented in an interpretable and transparent manner. To this end, the study developed ArchAI 2.0, a five-stage hierarchical system that models semantic and spatial relationships between two-dimensional floor plans and generated building facades. The objective was to conditionally generate exterior building image alternatives using floor plan-derived structural information. Following a pair-integrity screening, 657 valid floor plan–exterior photograph project pairs sourced from the ArchDaily platform were retained and partitioned at the project level into training ($n = 459$), validation ($n = 99$), and independent test ($n = 99$) sets. Plan-derived geometric information was extracted using a deterministic computer-vision parsing pipeline and subsequently processed into structural tokens and a depth-aware parametric facade skeleton. The resulting skeleton representation was used as conditioning guidance for a ControlNet integrated with Stable Diffusion 1.5. A Plan-Edge ControlNet trained and evaluated under the same protocol was used as the primary conditioned baseline, while an unconditional text-only Stable Diffusion 1.5 model provided a secondary reference. On the independent test set, the Skeleton and Plan-Edge models showed comparable performance, with no statistically significant differences in Chamfer Distance, Edge IoU, SSIM, or CLIP cosine similarity after Holm correction. Relative to the text-only baseline, Skeleton conditioning produced a modest but statistically significant improvement in Edge IoU (0.0611 vs. 0.0574; Holm-adjusted $p = 0.031$), whereas the text-only model achieved significantly higher CLIP similarity (0.2749 vs. 0.2647; Holm-adjusted $p < 0.001$). FID was treated as a descriptive dataset-level measure because of the limited test set size. A bounded human conditioning adherence evaluation further showed that 58 of 98 evaluable cases reproduced the requested two-story organization (59.18%; 58.59% across the full 99-case test set), with a mean absolute story error of 0.4184 stories. Opening-column correspondence was more limited: across 87 evaluable cases, bipartite matching at the primary threshold of $\tau = 0.08$ normalized facade width yielded a micro F1-score of 0.2970. An independent second-rater assessment on a predefined 25-case subset demonstrated high human annotation reliability, with 96.00% exact agreement and linearly weighted Cohen's $\kappa w = 0.9351$ for story-count ratings and an inter-rater opening-column matching F1-score of 0.7692. Rather than demonstrating uniform superiority or guaranteed realization of user-specified constraints, the findings indicate that the proposed depth-aware representation provides an alternative structured conditioning mechanism with a

Check for updates

Academic Editor: Yupeng Wang

Received: 1 July 2026
Revised: 13 September 2026
Accepted: 18 September 2026
Published: 27 October 2026

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

performance comparable to direct plan-edge guidance, while human evaluation reveals a clear distinction between representational editability and actual conditioning adherence. The deterministic rule engine complements this generative pipeline as an interpretable rule-based decision-support layer by expressing selected architectural relationships through explicit and reusable rules.

Keywords: explainable artificial intelligence (XAI); generative architectural design; floor plan parsing; ControlNet; deep learning

1. Introduction

Building facade design is influenced by environmental, functional, and user-centered parameters. Artificial intelligence-based methods are increasingly investigated as computational tools for architectural analysis, representation, and design support [1]. In this context, data-driven decision-support systems can provide a systematic mechanism for facade design by supporting the evaluation of multiple design variables [1,2]. Driven by advancements in computational power and deep learning, artificial intelligence has become increasingly relevant across architectural applications, including the extraction of patterns from design data, generative representation, and computational design support [3,4]. This development has encouraged artificial intelligence to be considered not merely as a tool for visual generation in architecture, but also as a computational approach for analyzing design decisions, representing spatial relationships, and supporting design-oriented decision processes.

Although deep generative models have demonstrated substantial potential for processing architectural data and synthesizing design alternatives, direct image-based conditioning can provide limited explicit control over how selected architectural relationships are represented during generation [5]. This limitation is particularly relevant to floor plan-to-facade generation because floor plans describe horizontal spatial organization, whereas facade images express vertical exterior composition, making the relationship between the two representations inherently indirect and one-to-many. To investigate an alternative structured representation for conditional facade generation, this study proposes a hierarchical decision-support system named ArchAI 2.0. Rather than directly translating plan pixels into facade pixels, ArchAI 2.0 introduces an intermediate geometric representation through sequential stages including deterministic plan parsing, structural token generation, and the construction of a parametric facade skeleton. In this framework, "depth-aware" refers specifically to a simplified relative depth encoding derived from normalized plan geometry and projected into a two-dimensional facade-oriented conditioning representation; it does not denote metric three-dimensional reconstruction. Selected design relationships are additionally expressed through a separate deterministic rule-based layer intended to support transparent decision support rather than to explain the internal reasoning of the generative model.

Although artificial intelligence applications are widely used across various architectural domains, including building design, floor plan generation, and facade synthesis, comparatively limited attention has been given to structured intermediate representations that connect floor plan-derived geometric information with conditional facade generation while retaining explicitly editable design parameters. Furthermore, generating a building facade from a floor plan is inherently a "one-to-many" problem, as a single plan can support multiple exterior design alternatives depending on designer intent, climatic conditions, material choices, and contextual requirements. Consequently, the proposed ArchAI 2.0 framework investigates conditional facade generation by integrating Stable Diffusion 1.5

and ControlNet with a floor plan-derived parametric facade skeleton and a separate deterministic rule-based interpretation layer. Rather than claiming to explain the internal decision-making process of the diffusion model, the rule-based component expresses selected architectural relationships through explicit and reusable "IF… THEN…" structures. Accordingly, the study investigates whether such a structured intermediate representation can serve as an editable conditioning interface for facade generation while explicitly distinguishing the editability of the representation from the extent to which its encoded constraints are realized in generated outputs. In this sense, ArchAI 2.0 is presented as an exploratory decision-support framework that combines structured conditional generation with transparent rule-based architectural relationships, rather than as a system guaranteeing geometrically faithful reconstruction of a source building.

2. Literature Review on Machine Learning and Deep Learning in Architecture

Artificial intelligence applications in architecture have evolved from rule-based and symbolic approaches, including shape grammars, graph-based representations, decision trees, and constraint-based systems, toward data-driven methods for architectural analysis and generation. Within this development, machine learning and deep learning have expanded the capacity of computational systems to learn representations from complex datasets and to support a broad range of analytical and generative tasks [6,7]. Convolutional neural networks (CNNs), in particular, have played an important role in image recognition, classification, feature extraction, and image-based representation [8–10], while recurrent architectures such as RNNs, LSTMs, and GRUs have extended learning-based methods to sequential and temporal data [11,12]. The growing use of generative models has subsequently shifted attention from classification and prediction toward the synthesis and conditional manipulation of design representations.

In architecture, these developments have supported applications across building-performance analysis, energy efficiency, facade design, plan generation, environmental design, and design decision support [4,13]. Deep learning has also been applied to architectural design and style classification [14,15], architectural drawing recognition and reconstruction [16,17], environmental and spatial visualization [18], and adaptive building applications [19]. Together, these studies illustrate the broad role of machine learning and deep learning within architectural research, extending from analytical and classification tasks to generative design processes. Hybrid approaches combining computational design rules with learning-based methods have further explored the relationship between explicit design knowledge and data-driven generation [20,21].

Within this broader development, generative artificial intelligence has become particularly relevant to architectural design. Generative adversarial networks (GANs) demonstrated the ability of learned models to synthesize architectural forms, facade configurations, and alternative visual representations from existing datasets [3,22,23]. Such approaches expanded the role of artificial intelligence from analysis and classification toward design exploration and image synthesis. However, recent advances in diffusion-based generative models have introduced an alternative framework for high-quality conditional image generation. Rombach et al. proposed latent diffusion models (LDMs), in which the diffusion process is performed within the latent representation of a pretrained autoencoder, reducing computational requirements while retaining high-resolution synthesis capabilities [24]. The introduction of cross-attention mechanisms also enables latent diffusion models to incorporate conditioning information such as text and other representations, providing the methodological foundation used by Stable Diffusion.

A particularly relevant development for architectural generation is the ability to introduce explicit spatial conditions into pretrained diffusion models. ControlNet, proposed by Zhang et al., extends pretrained text-to-image diffusion models by incorporating trainable conditional branches while keeping the underlying pretrained model parameters locked [5]. The framework supports spatial guidance through representations including edges, depth maps, segmentation maps, and other structural conditions. This capability is important for architectural applications because textual prompts alone may not adequately specify geometric organization, spatial composition, or the placement of architectural elements. ControlNet therefore provides a mechanism through which visual or geometric representations can participate directly in the conditioning process rather than relying exclusively on textual descriptions.

Recent architectural research has begun to investigate these diffusion-based and controllable generation strategies more directly. Stable Diffusion combined with LoRA and ControlNet has been investigated for building facade generation using structural conditions such as edge, depth, segmentation, and line-based representations, demonstrating the potential of explicit visual guidance to increase the controllability of architectural image synthesis [25]. Related architectural facade research has similarly combined Stable Diffusion with ControlNet-based edge and depth information to provide additional shape-related guidance during generation [26]. These developments indicate a shift from unconstrained architectural image synthesis toward generative workflows in which visual structure can be introduced explicitly as conditioning information.

Nevertheless, the relationship between floor plan organization and facade generation presents a particular challenge. A floor plan primarily represents horizontal spatial organization, including room arrangement, circulation, openings, and boundaries, whereas a facade represents vertical exterior composition. Consequently, the mapping between the two is not a deterministic reconstruction problem: a single floor plan may correspond to multiple plausible facade alternatives depending on architectural style, material selection, context, designer intent, and other design decisions. For this reason, the problem differs from both conventional architectural image synthesis and direct image-to-image transformation. The relevant question is not only whether a generative model can produce a plausible facade, but also how floor plan-derived spatial information can be represented as conditioning information for the generative process.

Against this background, the present study investigates a structured intermediate representation between floor plan geometry and conditional facade generation. ArchAI 2.0 converts plan-derived geometric information into structural tokens and subsequently into a parametric facade skeleton, which is used as conditioning guidance for a ControlNet integrated with Stable Diffusion 1.5. In this framework, "depth-aware" refers to a simplified relative depth encoding derived from normalized plan geometry and projected into a two-dimensional facade-oriented conditioning representation rather than to metric three-dimensional reconstruction. The Skeleton is therefore investigated as an alternative structured conditioning representation and is directly compared with simpler Plan-Edge conditioning under the same experimental protocol.

The framework also retains the decision-support perspective established in earlier artificial intelligence applications in architecture. However, explicit representation and generative realization are treated as separate issues. The ability to modify the parametric facade skeleton provides representational editability, but it does not guarantee that every encoded condition will be reproduced by the diffusion model; the degree of realization must instead be evaluated as conditioning adherence. Similarly, the deterministic rule-based component of ArchAI 2.0 expresses selected architectural relationships through explicit and reusable rules and is used as an auxiliary decision-support layer. It is not intended to

explain the internal latent operations or the decision-making process of Stable Diffusion or ControlNet. Accordingly, the contribution of the framework lies in investigating how structured floor plan-derived information can participate in conditional facade generation while retaining explicit architectural representations for design-oriented decision support.

3. Materials and Methods

The study developed ArchAI 2.0, a system incorporating a five-stage hierarchical decision-support mechanism designed to generate floor plan-conditioned exterior building images from 2D floor plans. The method comprises five sequential stages: (1) deterministic floor plan geometry parsing; (2) structural token extraction; (3) conversion of structural data into a depth-aware parametric facade representation; (4) use of this representation as guidance for a conditional diffusion model; and (5) the representation of selected predefined architectural relationships through explicit rule structures. Within this framework, "depth-aware" refers to a simplified relative depth encoding derived from normalized plan geometry and projected into a two-dimensional facade-oriented representation rather than to metric three-dimensional reconstruction. The generative component produces conditionally generated facade alternatives, whereas the auxiliary deterministic rule-based component separately expresses selected architectural relationships through explicit and reusable rules.

The fundamental concepts underlying the computational methods employed in this study are summarized below to facilitate understanding within the context of architectural design. The floor plan parsing stage used in the final experimental pipeline is a deterministic computer-vision geometry-extraction procedure and does not employ a learned segmentation network. Adaptive Gaussian thresholding, contour and polygon approximation, and line-based geometric detection are used to extract the principal exterior building perimeter and candidate exterior openings from the source floor plan drawings. The resulting geometry is normalized and subsequently processed by the structural token and parametric skeleton modules. The depth-aware parametric facade skeleton serves as an intermediate facade-oriented conditioning representation in which relative depth information derived from normalized plan geometry is encoded into a two-dimensional guidance image. Stable Diffusion 1.5, integrated with ControlNet, then uses this representation as a structural conditioning input for the generation of exterior building image alternatives [5]. Separately, an auxiliary deterministic rule engine expresses selected predefined architectural relationships through explicit "IF... THEN..." structures. These rules are author-defined heuristics rather than learned or statistically derived relationships and are intended to provide a transparent decision-support representation of selected architectural relationships; they do not explain the internal decision-making or latent operations of the diffusion model [27]. ChatGPT (OpenAI) was used as an auxiliary tool for debugging and minor corrections to existing code during implementation. All resulting modifications were reviewed and verified by the authors.

3.1. Dataset

The source collection contained 658 floor plan files and 658 exterior photograph files obtained from publicly accessible ArchDaily project pages published between January 2018 and December 2024 using the search keywords 'residential house', 'apartment building', and 'contemporary residence'. The web-scraping protocol gathered 658 projects encompassing a global geographic distribution across Europe (42%), Asia (30%), the Americas (22%), and other regions (6%). Typologically, the source collection comprised 65% single-family residential houses (428 projects) and 35% multi-family residential apartments (230 projects). Non-residential and adaptive-reuse projects were excluded to maintain typologi-

cal homogeneity. For projects containing multiple drawings or exterior images, a primary floor plan and a representative exterior photograph were selected on a project-by-project basis and manually verified as a project pair. All collected images were filtered to enforce a minimum spatial resolution of 512 × 512 pixels. Blurred drawings, low-contrast scans, hand-drawn sketches, and watermarked images were excluded. Near-duplicate projects were detected and removed using perceptual hashing (imagehash Hamming distance < 5). Because the source images are third-party copyrighted material, the raw image files are not publicly redistributed.

A final project-level integrity audit was performed before model training. Although 658 plan files and 658 exterior photograph files were present in the source directories, one floor plan (project ID 922) had no corresponding exterior photograph and one exterior photograph (project ID 920) had no corresponding floor plan. These unmatched records were excluded, resulting in 657 valid floor plan–exterior photograph project pairs used in the final experimental protocol. The 657 valid projects were partitioned at the project level into 459 training projects, 99 validation projects, and 99 independent test projects. The resulting split was stored as a fixed deterministic project-level partition, and no project ID overlapped across the three subsets.

Two conditional datasets were subsequently constructed from exactly the same project split. For the Plan-Edge condition, the conditioning image was derived exclusively from the corresponding floor plan using a deterministic preprocessing sequence consisting of contrast-limited adaptive histogram equalization (CLAHE), Gaussian smoothing, Canny edge detection (lower threshold = 30; upper threshold = 120), and one morphological dilation operation. For the Skeleton condition, each floor plan was processed through the deterministic floor plan geometry parsing, structural token extraction, and parametric facade skeleton pipeline to produce the corresponding structured conditioning representation. Both conditional datasets contained exactly the same 459 training, 99 validation, and 99 test project IDs. The paired exterior photographs were used as training targets but were not used in the construction of either conditioning representation, thereby preventing target-image information from entering the conditioning inputs.

No offline data augmentation was applied during the preparation of the final Plan-Edge or Skeleton datasets. Both datasets therefore retained one prepared conditioning–target pair per valid project within the fixed project-level partitions of 459 training, 99 validation, and 99 independent test projects. During ControlNet training only, random horizontal flipping was applied online with a probability of 0.5, and synchronously to each target exterior photograph and its corresponding conditioning image. This operation preserved the conditioning–target correspondence in image space; however, for architecturally asymmetric cases, horizontal mirroring cannot necessarily be assumed to preserve the semantic equivalence of left–right spatial relationships. This augmentation choice is therefore treated as a limitation of the training protocol rather than as an architecturally invariant transformation. No vertical flipping, rotation, scaling, translation, or shear augmentation was used. No augmentation was applied to the validation or test partitions. These partitions remained isolated from model training, and the independent test set was accessed only after training and checkpoint selection had been completed. The fixed project-level design ensured that the Plan-Edge and Skeleton conditioning strategies were trained and evaluated using identical project partitions under a common experimental protocol.

3.2. Proposed Method

ArchAI 2.0, proposed in this study, consists of five stages: deterministic floor plan geometry parsing, structural token extraction, depth-aware parametric skeleton generation, conditional facade synthesis, and, finally, an auxiliary deterministic rule-based decision-

support module. The stages form a hierarchical workflow in which intermediate representations derived from the floor plan are progressively transformed into structured geometric guidance for facade generation. The final rule-based stage is separate from the generative inference process and expresses selected predefined architectural relationships through explicit rule structures.

In the first stage, an RGB floor plan image is processed using a deterministic computer-vision geometry parser. The implemented parser applies adaptive Gaussian thresholding to separate drawing information from the background, followed by contour detection, polygonal approximation, and line- and perimeter-based geometric analysis to extract candidate representations of the principal exterior building perimeter and exterior opening locations. Extracted coordinates are normalized relative to the source image dimensions and stored in a structured semantic plan representation for subsequent token extraction and skeleton construction. The parser is entirely deterministic, and no learned FloorPlanUNet or other segmentation-network weights are used in the final experimental pipeline. Because the accuracy of this heuristic parsing stage directly affects the downstream representations, its performance was additionally assessed against manual annotations in a separate validation analysis described below.

In the second stage, the extracted geometric representation is processed by the deterministic TokenExtractor module. This module derives plan-based geometric descriptors such as wall lengths, opening locations, plan-level opening ratios, and overall mass proportions and stores the resulting structural tokens in JSON format. Vertical characteristics that cannot be recovered metrically from a two-dimensional floor plan are not treated as inferred physical dimensions; instead, selected vertical attributes are represented through predefined image space parameters used by the skeleton generator. Roof type is retained as a user-specified or default design parameter rather than being inferred from the source floor plan. The result of this stage is therefore a structured numerical representation of selected plan-derived geometric properties supplemented by explicitly defined design parameters required for subsequent skeleton construction.

In the third stage, the floor plan-derived geometric representation and selected structural tokens are transformed into a deterministic depth-aware parametric facade skeleton. In the locked benchmark configuration, the generator uses a front-view orthographic projection (view_angle = “front”), the bottom perimeter of the floor plan is designated as the reference facade direction (front_edge = “bottom”), and the number of above-ground stories is fixed at $N_f = 2$ for all benchmark projects. These settings are benchmark assumptions rather than project-specific quantities inferred from the source floor plan or paired exterior photograph. The active generator consumes the resulting floor-count setting together with roof-type and balcony-count parameters, while normalized wall, door, window, and perimeter geometry is obtained from the semantic plan representation. All geometric coordinates are normalized to the $[0, 1]$ range before being mapped to a 512×512-pixel conditioning canvas. Relative depth is derived from normalized plan geometry and represented through grayscale intensity within the two-dimensional facade-oriented projection. Accordingly, the resulting Skeleton is not a metrically scaled three-dimensional reconstruction and does not encode physical depth in real-world units. No information derived from the paired-reference exterior photograph is used during Skeleton construction. The implications of the fixed facade direction and story-count assumptions were subsequently examined through a manual metadata audit of the independent test set and post hoc sensitivity analyses using the frozen Skeleton checkpoint.

3.2.1. Parametric Skeleton Construction

The parametric skeleton was generated deterministically from the semantic representation of each floor plan. Coordinates describing the building outline, wall geometry, doors, and windows were first normalized by the width and height of the source floor plan image, producing coordinates in the interval $[0, 1]$. The horizontal extent of the resulting building profile was subsequently determined from the minimum and maximum x-coordinates of the available outline and opening geometry and mapped to the central 82% of a 512 × 512-pixel conditioning canvas.

In the final benchmark configuration, the bottom perimeter of the floor plan was defined as the reference facade direction (front_edge = "bottom"), and the generator operated using the frontal view (view_angle = "front"). This was a fixed benchmark assumption and was not inferred on a project-specific basis from the paired exterior photograph. Under this configuration, the representation is produced as an orthographic facade-oriented projection rather than through a perspective camera model. Relative depth is derived from the normalized vertical position of each plan outline segment. For a segment with mean normalized plan coordinate y_{avg}, relative depth d is calculated as

$$d = 1 - \frac{y_{avg} - y_{min}}{\max(0.01, y_{max} - y_{min})},$$

where y_{min} and y_{max} denote the minimum and maximum normalized vertical coordinates of the plan outline. Thus, d is bounded approximately within $[0, 1]$ for the evaluated outline geometry. Segments are ordered according to this relative depth before rasterization so that foreground elements can overwrite previously rendered background elements.

Depth information is encoded directly into the conditioning image using grayscale intensity. For each wall segment, the grayscale value I is computed as

$$I = clip(120 + 135d, 100, 255).$$

Consequently, the rendered wall surfaces span approximately 120–255 in the 8-bit conditioning image, with relative plan depth represented by differences in grayscale intensity. This encoding should therefore be understood as a normalized relative depth representation rather than metric physical depth.

Vertical organization is also defined in image space rather than physical units. For the 512 × 512-pixel canvas, an 8% vertical margin is used, giving a ground level at pixel 472. The floor height is calculated as

$$h_f = \min\left(\left\lfloor \frac{432}{N_f} \right\rfloor, 128\right),$$

where N_f denotes the number of represented above-ground stories. In the locked benchmark configuration, N_f was fixed at 2 for all projects rather than inferred individually from the source floor plans or paired exterior photographs, resulting in a floor height of 128 pixels. Horizontal slab boundaries are consequently positioned at integer multiples of this floor height above the ground level. The implementation does not assign a physical story height in meters; therefore, the Skeleton represents relative vertical organization rather than a metrically reconstructed elevation. The implications of this fixed two-story assumption were subsequently examined through the manual test set metadata audit and post hoc sensitivity analysis described below.

Openings are projected onto this facade-oriented representation using their normalized horizontal positions. Under the modern style configuration used in the final dataset preparation, door height is defined as 0.75 of the floor height and window height as 0.55 of the floor height, corresponding to 96 and 70 pixels, respectively, for the two-story configuration. Windows are positioned independently on each represented floor, whereas doors are

rendered at ground-floor level. These dimensions are proportional image space parameters and should not be interpreted as physical architectural dimensions.

Corners are obtained from the approximated building perimeter and represented as explicit vertical boundaries in the rasterized skeleton. Plan segments whose projected horizontal extent is less than three pixels collapse to vertical line markers in the frontal projection. Occlusion is handled using depth-ordered rasterization rather than geometric back-face culling: background segments are drawn first and can subsequently be overwritten by foreground segments. Minor deviations of less than 0.05 of the normalized plan dimension are treated as flat during the extraction of indentation depth.

The complete procedure is deterministic. No stochastic operation is used during semantic geometry-to-skeleton rasterization, and no random seed is therefore required for this stage. More importantly, skeleton construction receives only the floor plan image and its derived semantic representation; pixels, edges, or features from the paired-reference exterior photograph are not accessed. This separation prevents target-derived facade information from entering the conditioning representation.

The deterministic procedure used to generate the depth-aware parametric facade skeleton is summarized as follows:

- The floor plan image is parsed to extract candidate representations of the normalized building outline, wall geometry, door locations, and window locations.
- All geometric coordinates are normalized to the $[0, 1]$ interval using the width and height of the source floor plan image.
- The horizontal profile limits xmin and xmax are determined.
- The normalized building profile is mapped to the central 82% of the 512 × 512 conditioning canvas.
- The vertical layout parameters are defined as follows: $\text{margin} = \text{floor}(0.08 \times \text{S})$, $\text{ground_level} = \text{S} - \text{margin}$, and $\text{floor_height} = \min(\text{floor}((\text{S} - 2 \times \text{margin})/N_f), \text{floor}(0.25 \times \text{S}))$. For the locked benchmark configuration, $N_f = 2$.
- For each plan outline segment, the mean vertical plan coordinate is calculated and converted to relative depth using the normalized depth equation.
- Wall segments are sorted according to relative depth using a painter's-algorithm ordering.
- Each segment is projected onto the frontal conditioning canvas and encoded using grayscale intensity based on relative depth.
- Horizontal slab boundaries are drawn at intervals determined by floor height.
- Ground-floor doors are projected using a door-height ratio of 0.75 of floor height.
- Windows are projected on each represented floor using a window-height ratio of 0.55 of floor height.
- Structural outlines and the ground line are added.
- The resulting image is saved as a 512 × 512 RGB conditioning image.

In the fourth stage, the resulting Skeleton representation is used as the conditioning input to a ControlNet integrated with Stable Diffusion v1.5. To evaluate the proposed structured Skeleton representation relative to a simpler floor plan-derived conditioning strategy, a separate Plan-Edge ControlNet is trained under the same experimental protocol. In this comparison, the Plan-Edge condition is generated directly from the floor plan using CLAHE, Gaussian smoothing, Canny edge detection, and morphological dilation, whereas the Skeleton condition is generated through the semantic parsing, structural token extraction, and parametric skeleton pipeline. The two ControlNet models use the same Stable Diffusion v1.5 base model, identical project-level training, validation, and test partitions, and matched experimental settings, thereby isolating the conditioning representation as the principal experimental difference between the two conditioned configurations.

Alongside the core generative pipeline, ArchAI 2.0 includes an auxiliary logic-based rule engine that processes selected structural descriptors for interactive decision support. This module is separate from ControlNet training, final test set inference, image evaluation, and statistical analysis, and therefore does not affect the quantitative generation results reported in this study. The auxiliary layer maps predefined architectural relationships into explicit "IF... THEN..." rule structures and it is intended to provide a transparent representation of selected plan-derived descriptors and design parameters. The rule thresholds are author-defined heuristics rather than learned, statistically estimated, or empirically validated universal architectural relationships. Accordingly, this component is not presented as a predictive model or as an explanation of the internal latent decision-making process of Stable Diffusion or ControlNet.

3.2.2. ControlNet Training and Inference Protocol

To ensure a controlled comparison between conditioning representations, the Plan-Edge ControlNet and Skeleton ControlNet were trained using the same Stable Diffusion 1.5 base model (runwayml/stable-diffusion-v1-5) and were initialized from the same pretrained Canny ControlNet weights (lllyasviel/sd-controlnet-canny). During fine-tuning, the VAE, text encoder, and UNet parameters were frozen, and only the ControlNet parameters were updated.

Both models were trained on the same 459-project training partition for 3000 optimization steps at a resolution of 512 × 512 pixels. AdamW was used with a learning rate of 1×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. A cosine learning-rate schedule with 300 warm-up steps was applied. The per-device batch size was 4, with four gradient accumulation steps, yielding an effective batch size of 16. Training used bfloat16 mixed precision and gradient clipping with a maximum norm of 1.0. The optimization objective was the standard diffusion mean-squared error between the predicted and sampled noise.

No offline augmentation was applied when constructing either the Plan-Edge or Skeleton datasets. During model training only, a random horizontal flip with a probability of 0.5 was applied synchronously to each target exterior photograph and its corresponding conditioning image. This operation preserved the conditioning–target correspondence in image space; however, horizontal mirroring cannot necessarily be assumed to preserve the architectural semantics of left–right relationships in asymmetric buildings. The augmentation is therefore treated as a limitation of the training protocol rather than as an architecturally invariant transformation. No augmentation was applied during validation or testing. Training samples were shuffled, whereas validation samples were processed without shuffling.

Model selection was based exclusively on the 99-project validation partition. Validation was performed every 250 optimization steps using a deterministic noise generator initialized with seed 42, while regular checkpoints were saved every 500 steps. The checkpoint yielding the lowest validation diffusion loss was retained as the best model. For both conditioning strategies, the selected checkpoint occurred at step 1750, with validation losses of 0.101227 for Plan-Edge ControlNet and 0.101224 for Skeleton ControlNet. The training continued to the full 3000 steps; early stopping was not used. The independent 99-project test partition was not accessed during training, validation, or checkpoint selection.

Final inference was performed on the locked 99-project test partition using three configurations: text-only Stable Diffusion 1.5, Plan-Edge ControlNet, and Skeleton ControlNet. All configurations used the UniPCMultistepScheduler with 30 inference steps, a classifier-free guidance scale of 7.5, and an output resolution of 512 × 512 pixels. For the two ControlNet configurations, the conditioning scale was fixed at 1.0. The same positive prompt—"contemporary residential building exterior, architectural photography,

natural daylight, realistic materials, clear facade composition"—and the same negative prompt—"blurry, low quality, distorted geometry, deformed windows, deformed doors, text, watermark"—were used for all three conditions.

To maintain paired stochastic conditions while avoiding identical generations across different projects, each test project was assigned a deterministic seed according to $s_i = 42 + ID_i$, where ID_i denotes the integer project identifier. The same project-specific seed was used for the text-only, Plan-Edge, and Skeleton configurations of a given project. Consequently, differences among the three outputs for each test case arise from the conditioning configuration rather than from differences in prompt, inference schedule, or random seed.

Figure 1 summarizes the core architecture of the proposed ArchAI 2.0 framework together with the controlled benchmark protocol used to compare the Skeleton-conditioned model with the Plan-Edge ControlNet baseline and the text-only Stable Diffusion 1.5 reference.

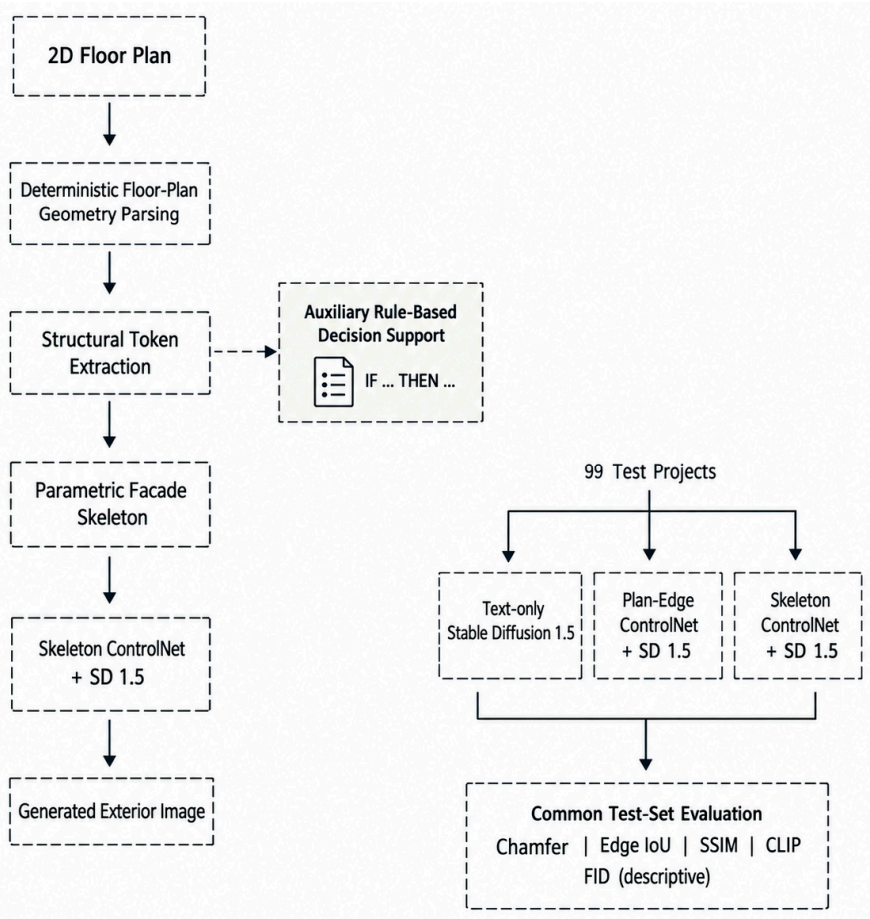


Figure 1. Overview of the core ArchAI 2.0 generative framework and controlled benchmark comparison. The proposed pathway transforms the input floor plan through deterministic floor plan geometry parsing, structural token extraction, and parametric facade skeleton generation before conditional exterior image synthesis using a Skeleton ControlNet integrated with Stable Diffusion 1.5. The auxiliary deterministic rule-based decision-support component operates separately from the generative inference and quantitative evaluation pipeline. The Plan-Edge baseline derives its conditioning representation directly from the same floor plan, whereas the text-only Stable Diffusion 1.5 configuration provides a reference without plan-derived structural conditioning. All three generation configurations are evaluated on the same independent 99-project test partition using matched prompts, project-specific random seeds, and inference settings. Chamfer Distance, Edge IoU, SSIM, and CLIP similarity are evaluated at the project level, whereas FID is reported only as a descriptive dataset-level measure.

3.3. Evaluation Protocol

The final evaluation was conducted on the locked independent test set of 99 projects using three generation conditions: text-only Stable Diffusion 1.5, Plan-Edge ControlNet, and Skeleton ControlNet (ArchAI 2.0). For the paired-reference image evaluation, each generated exterior image was compared with the corresponding project reference exterior photograph using a common preprocessing and metric protocol. All generated images were evaluated at 512 × 512 pixels, and reference exterior photographs were symmetrically padded to a square canvas when necessary and resized to the same resolution using Lanczos interpolation. Because these reference images are project photographs rather than geometrically registered orthographic elevations, the resulting image-based metrics are interpreted as paired-reference similarity measures rather than direct measures of floor plan or conditioning fidelity.

Chamfer Distance and Edge IoU were computed between edge maps independently extracted from the generated exterior image and the paired-reference exterior photograph. For both images, RGB data were converted to grayscale, smoothed using a 3 × 3 Gaussian kernel, and processed using Canny edge detection with lower and upper thresholds of 50 and 150, respectively. No conditioning edge map, parametric skeleton image, or target-derived control signal was reused in the metric computation. Edge IoU was calculated as the ratio between the intersection and union of the two binary edge maps. Chamfer Distance was calculated symmetrically as the mean nearest-edge distance from generated image edge pixels to reference image edge pixels and from reference image edge pixels to generated image edge pixels, expressed in pixels.

Because the reference data consist of perspective exterior photographs rather than orthographic architectural elevations, Chamfer Distance and Edge IoU are not interpreted as direct measures of floor plan geometric accuracy. Instead, they are treated as paired-reference edge-correspondence proxies that quantify similarity between the edge structure of the generated exterior image and the corresponding reference photograph. This distinction is necessary because floor plans and perspective exterior photographs represent different geometric projections and cannot be assumed to be directly commensurable in image space.

SSIM was calculated conventionally between the generated RGB exterior image and the paired-reference RGB exterior photograph, using an 11 × 11 Gaussian window with $\sigma = 1.5$ and a data range of 255. Therefore, SSIM in this study represents image-level structural similarity to the paired-reference photograph and does not compare the parametric skeleton with the generated image.

CLIP cosine similarity was calculated between each generated exterior image and the fixed text prompt used during inference, using the pretrained CLIP ViT-B/32 model. FID was calculated at the dataset level between the 99 generated images and the 99 reference exterior photographs using 2048-dimensional Inception-v3 features. Because the test set size was limited to $n = 99$, FID was retained only as a descriptive distribution-level indicator and was not used for inferential statistical testing.

3.3.1. Conditioning Adherence Evaluation

In addition to the paired-reference image metrics, a bounded human conditioning-adherence evaluation was conducted to assess more directly whether selected structural information encoded in the Skeleton condition was realized in the generated exterior images. This evaluation compared the frozen Skeleton conditioning representation with its corresponding generated output and did not use the paired third-party exterior photograph as a geometric target. The evaluation was conducted on the complete predefined 99-project independent test set and it considered two observable attributes: represented story organization and horizontal opening-column correspondence.

For story organization, the evaluator recorded the number of visually distinguishable above-ground stories in each generated exterior image. Cases in which the story organization could not be resolved because of severe visual occlusion or ambiguity were marked as unresolvable rather than assigned an inferred value. The observed story count was compared with the story organization requested by the corresponding Skeleton condition. Adherence was summarized using the number and proportion of exact matches together with the absolute story-count error.

For opening correspondence, visually identifiable opening columns in the generated facade were annotated and compared with the horizontal opening-column positions encoded in the corresponding Skeleton condition. Horizontal positions were represented in normalized facade width coordinates. Correspondence was determined through one-to-one bipartite matching subject to a maximum horizontal centroid-distance threshold, with $\tau = 0.08$ of normalized facade width defined as the primary threshold. Matches within this tolerance were counted as true positives; unmatched generated opening columns were counted as false positives, and unmatched Skeleton opening columns as false negatives. Micro-averaged precision, recall, and F1-score were then calculated across all evaluable cases. To examine sensitivity to the matching tolerance, the same analysis was additionally repeated at $\tau = 0.05$ and $\tau = 0.10$. Cases in which opening boundaries could not be evaluated reliably because of heavy texture, occlusion, or visually ambiguous facade organization were recorded as unresolvable and excluded from the evaluable-case opening analysis.

To assess annotation reliability, a second independent evaluator repeated the same story-count and opening-column annotation procedure on a predefined 25-project subset selected independently of the resulting adherence scores. Story-count agreement between evaluators was assessed using exact agreement and linearly weighted Cohen's kappa. For opening annotations, inter-rater correspondence was assessed using the same primary matching tolerance ($\tau = 0.08$), with precision, recall, and F1-score calculated between the two sets of annotated opening-column positions. Differences in the annotated left and right facade boundaries were also recorded as supplementary indicators of annotation variability. The second-rater analysis was used to evaluate the reliability of the human annotation procedure and was not interpreted as an additional measure of model conditioning adherence.

3.3.2. Test Set Metadata Audit and Post Hoc Sensitivity Analysis

To examine the implications of the fixed facade direction and story-count assumptions used in the locked Skeleton benchmark, a post hoc manual metadata audit was conducted on all 99 projects in the independent test set. This audit was performed after completion of the primary benchmark and did not alter the original training, validation, checkpoint selection, or test generation protocol. Each project was manually reviewed using the available project-level plan and exterior image information to characterize two attributes relevant to Skeleton construction: the facade direction represented by the paired exterior photograph relative to the floor plan and the number of visible above-ground stories.

For facade direction, each project was categorized as BOTTOM, TOP, LEFT, RIGHT, or AMBIGUOUS according to the side of the floor plan judged to correspond most plausibly to the facade represented in the paired exterior photograph. Cases in which a sufficiently reliable correspondence could not be established from the available project information were retained as AMBIGUOUS rather than being assigned a forced orientation. These manually reviewed categories were treated as project-level metadata for sensitivity analysis rather than as geometrically registered ground-truth orientation labels.

Visible above-ground story count was independently recorded from the available project-level exterior information. Cases for which a reliable story count could not be established were categorized as UNKNOWN. As with facade direction, these annota-

tions were used as manually reviewed project-level metadata and were not treated as metric geometric ground truth. The audit was used to quantify how frequently the fixed benchmark assumptions of front_edge = "bottom" and $N_f = 2$ agreed with the available project-level information.

Following the audit, post hoc sensitivity analyses were performed using the frozen Skeleton ControlNet checkpoint selected in the original experiment. No retraining, additional fine-tuning, or checkpoint reselection was performed. For projects with a resolved non-bottom facade direction, the Skeleton conditioning representation was regenerated after replacing the fixed bottom reference edge with the manually reviewed project-specific direction. For projects with a known above-ground story count different from two, the Skeleton was regenerated using the manually reviewed story count. A combined intervention was additionally evaluated for projects in which both a resolved non-bottom facade direction and a non-two-story count were available. All other inference settings, including the trained checkpoint, prompts, project-specific random seeds, scheduler, guidance settings, and output resolution, were retained from the locked benchmark protocol.

These analyses were designed as post hoc sensitivity interventions rather than as a retrained benchmark or a new project-specific generation model. Their purpose was to examine whether correcting the two fixed Skeleton-construction assumptions produced systematic changes in the frozen model outputs. Comparisons were therefore restricted to the eligible cases for each intervention and were interpreted as sensitivity evidence concerning the fixed orientation and story-count assumptions rather than as estimates of general model performance. Direct realization of the encoded structural conditions was assessed separately through the conditioning adherence evaluation described in Section 3.3.1.

3.3.3. Representational Editability Protocol

A post hoc representational editability analysis was conducted to examine whether the two Skeleton-construction parameters directly implicated by the fixed benchmark assumptions—represented floor count and facade direction—could be modified at inference time without retraining the Skeleton ControlNet. Three projects were selected from the independent test set using a predefined, performance-independent percentile procedure based on the ordered test projects with a non-zero plan-level window ratio. The selected projects corresponded to the 25th, 50th, and 75th percentile positions (Projects 245, 463, and 710, respectively). This selection was performed independently of the generation results produced by the editability analysis.

For each project, two parameter families were evaluated independently. Represented floor count was varied across floor_count = 1, 2, and 3 while front_edge = "bottom" was held fixed. Facade direction was independently varied across front_edge = "bottom", "left", and "right" while floor_count = 2 was held fixed. Only one parameter was varied at a time, and all other Skeleton-generation and inference settings were retained from the locked benchmark configuration. The same project-specific deterministic seed was reinitialized for every variant within each project, thereby providing identical initial noise for the corresponding comparisons.

Six variants were generated for each project, producing 18 generated exterior images and their 18 corresponding Skeleton conditioning representations. All variants used the same frozen Skeleton ControlNet checkpoint employed in the primary experiment; no retraining, fine-tuning, checkpoint reselection, prompt modification, or adjustment of the ControlNet conditioning scale was performed. The analysis was qualitative and was designed specifically to examine representational editability and the response of the frozen model to modified conditioning representations. It was not designed to establish precise out-

put controllability or guaranteed realization of the edited parameters, which was assessed separately through the conditioning adherence analysis described in Section 3.3.1.

3.3.4. Deterministic Parser Validation

Because the deterministic floor plan parser provides the geometric information used by the downstream structural token and Skeleton-generation stages, a separate manual validation was conducted to characterize its extraction performance. A stratified sample of 40 projects was selected from the fixed project-level dataset partitions using a deterministic random seed of 42, comprising 28 training, six validation, and six independent test projects. For each sampled floor plan, manual reference annotations were prepared for the principal exterior perimeter, exterior windows, and exterior doors. These annotations were used only for post hoc parser validation and did not modify the dataset partitions, parser configuration, ControlNet training, checkpoint selection, or final generation benchmark.

Parser performance was evaluated separately for perimeter geometry and exterior-opening detection. Correspondence of the extracted principal exterior perimeter with the manual annotation was quantified using perimeter Intersection over Union (IoU) and boundary Chamfer Distance in pixels. Exterior window and exterior door detection performance was evaluated using micro-averaged precision, recall, and F1-score across the 40-project validation sample. Because the parser is a deterministic heuristic geometry-extraction procedure rather than a learned segmentation or object-detection model, this analysis was intended to characterize the quality and limitations of the geometric information propagated to the downstream ArchAI 2.0 stages rather than to constitute an independently optimized parser benchmark.

4. Results

The performance of ArchAI 2.0 was evaluated on the locked independent test set of 99 projects using complementary paired-reference image metrics and direct conditioning adherence measures. Paired-reference edge correspondence was assessed using Chamfer Distance and Edge IoU, whereas image-level structural and semantic similarity was evaluated using the Structural Similarity Index Measure (SSIM) [28], Fréchet Inception Distance (FID) [29], and CLIP cosine similarity [30]. In addition, a bounded human evaluation assessed the realization of selected structural information encoded in the Skeleton condition through story organization and opening-column correspondence. Because floor plan-to-facade generation is inherently a one-to-many problem and the paired project photographs are not geometrically registered architectural elevations, the image-based metrics are interpreted as complementary paired-reference proxies rather than as direct measures of floor plan fidelity or overall architectural quality.

Table 1 presents the quantitative results for the three controlled experimental conditions: the text-only Stable Diffusion 1.5 reference, the Plan-Edge ControlNet baseline, and the proposed Skeleton ControlNet configuration of ArchAI 2.0. All three conditions were evaluated on the same 99 independent test projects using matched prompts, project-specific deterministic seeds, inference settings, and output resolution. Sample-based metrics are reported as mean ± 95% confidence-interval half-width, whereas FID is reported as a dataset-level descriptive point estimate.

Table 1. Quantitative performance of the three controlled generation conditions on the independent test set (n= 99).

Configuration	Edge IoU ↑	Chamfer Distance (px) ↓	SSIM ↑	FID ↓	CLIP Cosine Similarity ↑
Text-only SD 1.5	0.0574 ± 0.0037	14.4312 ± 1.2185	0.1298 ± 0.0128	209.0824	0.2749 ± 0.0021
Plan-Edge ControlNet + SD 1.5	0.0609 ± 0.0044	15.0479 ± 1.3129	0.1276 ± 0.0128	178.7036	0.2646 ± 0.0026
Skeleton ControlNet + SD 1.5 (ArchAI 2.0)	0.0611 ± 0.0045	15.1463 ± 1.2967	0.1296 ± 0.0132	180.4894	0.2647 ± 0.0024

Note: Values for Edge IoU, Chamfer Distance, SSIM, and CLIP cosine similarity are reported as mean ± 95% confidence-interval half-width across the 99 independent test projects. FID is calculated at the dataset level and is reported descriptively; no inferential statistical test was applied to FID. ↑ indicates higher is better; ↓ indicates lower is better.

Figure 2 provides a qualitative complement to the quantitative evaluation by showing outputs from the three generation configurations for two representative test cases and one limitation case. Visual inspection indicates that structural conditioning alters facade organization and massing relative to the text-only condition; however, the Skeleton ControlNet does not produce uniformly superior outputs to the Plan-Edge ControlNet. In particular, Project 287 illustrates a limitation of the simplified front-projection Skeleton and does not show an evident qualitative advantage over the Plan-Edge condition. These examples are therefore presented as qualitative illustrations of model behavior rather than as evidence of general superiority of any individual configuration.



Figure 2. Qualitative comparison of the final generation configurations across three projects from the independent test set. Column (**A**) shows the depth-encoded parametric facade skeleton condition

derived from the corresponding floor plan. Columns (**B**–**D**) show the generated exterior building images obtained using text-only Stable Diffusion 1.5, Plan-Edge ControlNet, and Skeleton ControlNet, respectively. Projects 607 and 610 illustrate representative test cases, whereas Project 287 is included as a limitation case in which the simplified front-projection Skeleton does not show an evident qualitative advantage over the Plan-Edge condition. The examples are provided for qualitative illustration and are not used as evidence of source-plan fidelity. The original floor plans and paired-reference exterior photographs are not reproduced because they originate from third-party copyrighted project sources.

4.1. Quantitative and Statistical Analysis

The quantitative results indicate that the three generation conditions exhibit different performance profiles rather than a uniform ranking across all evaluation metrics. The proposed Skeleton ControlNet configuration achieved the highest mean Edge IoU (0.0611 ± 0.0045), compared with 0.0609 ± 0.0044 for the Plan-Edge ControlNet and 0.0574 ± 0.0037 for the text-only SD 1.5 reference. In contrast, the text-only condition yielded the lowest mean Chamfer Distance (14.4312 ± 1.2185 px), followed by the Plan-Edge ControlNet (15.0479 ± 1.3129 px) and Skeleton ControlNet (15.1463 ± 1.2967 px). These results suggest that the conditioning strategies affect the measured paired-reference edge-correspondence proxies differently and therefore do not support claims of uniform geometric superiority.

The non-zero Edge IoU observed for the text-only condition should not be interpreted as evidence of floor plan compliance. Because Edge IoU is computed against the paired-reference exterior photograph rather than against the source floor plan or conditioning representation, generic exterior image edges such as rooflines, openings, and building boundaries may produce partial overlap even without explicit floor plan conditioning.

For the image-level metrics, SSIM values were similar across the three conditions, with means of 0.1298 ± 0.0128 for the text-only reference, 0.1276 ± 0.0128 for the Plan-Edge ControlNet, and 0.1296 ± 0.0132 for the Skeleton ControlNet. The text-only condition achieved the highest CLIP cosine similarity (0.2749 ± 0.0021), whereas the Plan-Edge and Skeleton ControlNet conditions yielded 0.2646 ± 0.0026 and 0.2647 ± 0.0024, respectively. At the dataset level, both conditioned models produced lower descriptive FID values than the text-only condition: 178.7036 for Plan-Edge ControlNet and 180.4894 for Skeleton ControlNet, compared with 209.0824 for text-only SD 1.5. Given the test set size ($n = 99$), FID is interpreted here only as a descriptive distribution-level indicator and was excluded from inferential statistical testing.

Because the same 99 independent test projects were evaluated under all three conditions, statistical comparisons were performed using paired analyses. The normality of the paired differences was assessed using the Shapiro–Wilk test. A paired two-tailed t-test was used when the normality assumption was not rejected; otherwise, a Wilcoxon signed-rank test was applied. Bootstrap 95% confidence intervals were calculated for the mean paired differences, and Holm adjustment was used to control for multiple comparisons within each comparison family.

The primary statistical comparison evaluated the proposed Skeleton ControlNet against the Plan-Edge ControlNet baseline. No statistically significant differences were observed between these two conditioned models for any of the four sample-level metrics after Holm adjustment. For Chamfer Distance, the mean paired difference (Skeleton − Plan-Edge) was 0.0984 px (bootstrap 95% CI $[-0.5011, 0.6954]$; Wilcoxon $p = 0.7376$; Holm-adjusted $p = 1.000$; matched-pairs rank-biserial effect size = 0.0388). For Edge IoU, the corresponding difference was 0.00025 (95% CI $[-0.00136, 0.00179]$; Wilcoxon $p = 0.8587$; Holm-adjusted $p = 1.000$; rank-biserial effect size = 0.0206). Similarly, no significant differences were observed for SSIM (mean difference = 0.00198; 95% CI $[-0.00085, 0.00482]$; Holm-adjusted $p = 1.000$) or CLIP cosine similarity (mean difference = 0.00011; 95% CI $[-0.00207, 0.00223]$;

Holm-adjusted $p = 1.000$). Thus, although the Skeleton ControlNet achieved slightly higher numerical means for Edge IoU, SSIM, and CLIP cosine similarity, the paired analysis does not provide evidence of statistical superiority over the simpler Plan-Edge ControlNet baseline.

In the secondary comparison with the text-only SD 1.5 reference, the Skeleton ControlNet achieved a significantly higher Edge IoU. The mean paired difference was 0.00376 (bootstrap 95% CI [0.00103, 0.00661]), and the paired two-tailed t-test yielded $p = 0.0103$ and a Holm-adjusted $p = 0.0310$, with a small effect size (Cohen's d_z = 0.2628). However, no statistically significant differences were detected for Chamfer Distance (Holm-adjusted $p = 0.4204$) or SSIM (Holm-adjusted $p = 0.6984$). In contrast, CLIP cosine similarity was significantly lower for the Skeleton ControlNet than for the text-only condition, with a mean paired difference of −0.01025 (bootstrap 95% CI [−0.01321, −0.00730]; Holm-adjusted $p < 0.001$; Cohen's d_z = −0.6842). This indicates that the improvement in pixel-level edge overlap was accompanied by reduced text-image semantic similarity under the present evaluation protocol.

The exploratory comparison between the Plan-Edge ControlNet and the text-only reference showed a similar pattern. Plan-Edge ControlNet produced a numerically higher Edge IoU, with a mean paired difference of 0.00351, although the difference did not remain statistically significant after Holm adjustment (adjusted $p = 0.0718$). No statistically significant differences were observed for Chamfer Distance or SSIM. CLIP cosine similarity, however, was significantly lower for the Plan-Edge ControlNet than for the text-only condition (Holm-adjusted $p < 0.001$; Cohen's d_z = −0.6405).

Overall, the controlled evaluation does not demonstrate uniform superiority of the proposed Skeleton ControlNet across the selected metrics. Instead, the results indicate a more specific effect: Skeleton-based conditioning provides a small but statistically detectable improvement in Edge IoU relative to text-only generation, while its performance is statistically comparable to the Plan-Edge ControlNet baseline across the evaluated sample-level metrics. The accompanying reduction in CLIP cosine similarity further indicates that the conditioned configurations do not improve all evaluation dimensions simultaneously. Accordingly, the contribution of the Skeleton representation should not be interpreted as universal metric improvement over simpler Plan-Edge conditioning. Rather, its proposed value lies in providing an explicit structured and editable intermediate conditioning representation derived from floor plan geometry. Whether the structural information encoded in this representation is actually realized in the generated outputs is a separate question and is examined directly through the conditioning adherence evaluation presented in the following subsection.

4.2. Conditioning Adherence Results

The conditioning adherence evaluation provided a direct assessment of whether selected structural information encoded in the parametric facade skeleton was realized in the corresponding Skeleton ControlNet outputs. Unlike the paired-reference metrics reported above, this analysis compared the conditioning representation directly with the generated exterior image and did not use the paired project exterior photograph as a geometric target.

For story organization, 98 of the 99 test cases were evaluable, while one case was classified as unresolvable because severe visual occlusion prevented a reliable story-count assessment. Among the 98 evaluable cases, 58 generated images reproduced the requested two-story organization, corresponding to an exact-match rate of 59.18%. When expressed relative to the complete 99-project test set, including the single unresolvable case, the exact-match rate was 58.59%. The remaining 40 evaluable cases differed from the requested story count. The generated outputs comprised 30 one-story, 58 two-story, nine three-story, and one four-story cases, with one additional unresolvable case. The mean absolute story-count

error across evaluable cases was 0.4184 stories, with a median error of 0. These results indicate moderate, but not deterministic, realization of the story organization encoded in the conditioning representation.

Opening-column correspondence was more limited. Of the 99 test cases, 87 were considered evaluable for opening correspondence, whereas 12 were classified as unresolvable because heavy texture, occlusion, or ambiguous opening boundaries prevented reliable annotation. Across the 87 evaluable cases, the Skeleton conditions contained 180 opening columns and the corresponding generated facades contained 224 annotated opening columns. At the predefined primary matching threshold of $\tau = 0.08$ normalized facade width, one-to-one bipartite matching yielded 60 true positives, 164 false positives, and 120 false negatives, corresponding to a micro-averaged precision of 0.2679, recall of 0.3333, and F1-score of 0.2970. The mean matched-centroid error was 0.0394 normalized facade width, with a median of 0.0393.

The threshold-sensitivity analysis produced the expected variation in correspondence. At the stricter threshold of $\tau = 0.05$, 40 true positives, 184 false positives, and 140 false negatives yielded a precision of 0.1786, recall of 0.2222, and F1-score of 0.1980. At the more permissive threshold of $\tau = 0.10$, 73 true positives, 151 false positives, and 107 false negatives yielded a precision of 0.3259, recall of 0.4056, and F1-score of 0.3614. Thus, although correspondence increases as the matching tolerance is relaxed, the opening-column analysis consistently indicates limited realization of the horizontal opening structure encoded in the Skeleton condition.

The independent second-rater analysis indicated that the observed adherence limitations cannot be attributed solely to poor reproducibility of the human annotation procedure. On the predefined 25-project subset, the two evaluators agreed exactly on story count in 24 of 25 cases (96.00%), with a linearly weighted Cohen's κ of 0.9351. For opening-column annotations at the primary threshold of $\tau = 0.08$, inter-rater matching yielded 40 true positives, 12 false positives, and 12 false negatives, corresponding to precision, recall, and F1-score values of 0.7692. The mean normalized centroid disagreement between matched opening annotations was 0.0143. These values characterize inter-rater annotation reliability rather than model conditioning adherence.

Table 2 summarizes the principal conditioning adherence results together with the second-rater reliability assessment. The results show that story organization was realized more consistently than opening-column correspondence, while the substantially higher inter-rater agreement indicates that the limited opening correspondence primarily reflects model conditioning adherence rather than annotation inconsistency.

Table 2. Summary of conditioning adherence and annotation-reliability results for the Skeleton ControlNet.

Evaluation	Cases/Subset	Result
Story exact match	98/99	58/98 (59.18%)
Mean absolute story-count error	98/99	0.4184 stories
Opening correspondence, $\tau = 0.05$	87/99	F1 = 0.1980
Opening correspondence, $\tau = 0.08$ (primary)	87/99	F1 = 0.2970
Opening correspondence, $\tau = 0.10$	87/99	F1 = 0.3614
Second-rater story agreement	25	96.00%; weighted $\kappa = 0.9351$
Second-rater opening agreement, $\tau = 0.08$	25	F1 = 0.7692

Note: Model conditioning adherence results and inter-rater reliability results represent different analyses and should not be interpreted interchangeably. Opening-column F1-scores are based on one-to-one matching at the indicated normalized facade width thresholds.

Taken together, the human evaluation reveals an important distinction between representational editability and conditioning adherence. The parametric facade skeleton provides an explicit representation in which selected structural attributes can be specified or modified, but the diffusion model does not reliably reproduce every encoded constraint in the generated exterior image. Story organization showed moderate realization, whereas horizontal opening-column correspondence was substantially more limited. The Skeleton should therefore be interpreted as a structured conditioning interface rather than as a mechanism for guaranteed or geometrically precise constraint realization.

4.3. Interpretable Rule-Based Decision-Support Layer

In this study, an auxiliary deterministic logic-based rule engine was implemented as an interpretable decision-support layer alongside the generative pipeline. The module evaluates selected structural descriptors and maps predefined relationships into explicit "IF... THEN..." rule structures. It is not invoked during ControlNet training, final test set inference, image-metric evaluation, or statistical analysis and therefore does not contribute to the quantitative performance results reported in Section 4.1. Importantly, the rule engine does not constitute a learned surrogate of the diffusion model and does not reconstruct or explain the internal decision-making process of Stable Diffusion or ControlNet. Its role is limited to providing a transparent and reproducible representation of selected author-defined architectural relationships for interactive decision support.

The numerical thresholds used in the auxiliary rule-based decision-support layer were defined heuristically during system development as explicit operating parameters rather than being statistically learned or optimized from the study dataset. The window-ratio threshold of 0.40 and the aspect-ratio threshold of 1.30 are dimensionless. Projection depth and indentation depth are derived from normalized plan space depth values and multiplied by a factor of 10.0 by the TokenExtractor; therefore, the threshold value of 1.0 corresponds to a normalized relative plan depth of 0.10 and does not represent one meter or any other physical distance. These thresholds should consequently be interpreted as prototype decision parameters used to demonstrate deterministic rule processing rather than as empirically validated architectural standards. Their calibration and validation using larger architectural datasets and expert-based evaluation remain subjects for future work.

The rules implemented in the logic-based rule engine are presented below.

Window Ratio (Rule 1)

IF the window ratio is ≥0.40, THEN the rule engine associates the configuration with a higher-transparency facade interpretation characterized by larger glazed surfaces.

Balance of Solids and Voids (Rule 2)

IF the window ratio is <0.40, THEN the rule engine associates the configuration with a more opaque facade interpretation characterized by a greater proportion of solid surfaces.

Projection Depth (Rule 3)

IF the scaled projection-depth descriptor is >1.0, THEN the rule engine identifies pronounced projection or volumetric articulation as a corresponding facade-design interpretation.

Indentation Depth (Rule 4)

IF the scaled indentation-depth descriptor is >1.0, THEN the rule engine identifies recessed surfaces or entrance-niche characteristics as a corresponding facade-design interpretation.

Roof Typology (Rule 5)

IF roof_type = flat, THEN the rule engine associates the configuration with horizontal roof-edge characteristics and a simplified upper-boundary geometry.

IF roof_type = gable, THEN the rule engine associates the configuration with pitched-roof geometry and a triangular gable expression. In the final implementation, roof_type is treated as a user-specified or default design attribute rather than as a parameter automatically inferred from floor plan geometry.

Mass Ratio (Rule 6)

IF the aspect ratio is ≥1.30, THEN the rule engine associates the configuration with a horizontally emphasized facade interpretation, including linear facade organization and horizontally distributed openings.

These rules describe the explicit deterministic logic implemented by the auxiliary decision-support layer rather than causal explanations of individual diffusion-model outputs. In the present study, the rule layer was not quantitatively validated as a predictor of the generated exterior images, nor were its thresholds statistically calibrated against architectural performance outcomes or expert judgments. Its contribution is therefore limited to providing a transparent and reproducible implementation of selected author-defined design relationships within the broader ArchAI 2.0 decision-support framework. Accordingly, the rule-based component should be interpreted as a prototype decision-support mechanism rather than as a validated architectural inference model or an explanation of the internal operation of the generative model. The evaluation of the architectural validity, practical usefulness, and threshold sensitivity of these rules through expert studies and empirical calibration remains necessary in future work.

4.4. Test Set Metadata Audit and Post Hoc Sensitivity Analysis

The post hoc metadata audit demonstrated that the fixed facade direction and story-count assumptions used in the locked benchmark did not correspond uniformly to the available project-level information. For facade direction, 43 of the 99 test projects were categorized as BOTTOM, eight as RIGHT, four as LEFT, one as TOP, and 43 as AMBIGUOUS. Among the 56 projects for which a facade direction could be resolved, 43 (76.8%) were consistent with the benchmark assumption of front_edge = "bottom", whereas 13 (23.2%) indicated a different reference direction. The high number of ambiguous cases further indicates that a unique correspondence between a floor plan edge and the facade represented in a project photograph cannot always be established reliably from the available source material.

The story-count audit similarly showed substantial variation relative to the fixed two-story benchmark configuration. Of the 99 projects, 92 had a resolvable visible above-ground story count and seven were categorized as UNKNOWN. Among the 92 known cases, 36 were classified as one story, 38 as two stories, 11 as three stories, five as four stories, one as seven stories, and one as eight or more stories. Thus, only 38 of the 92 projects with known story information (41.3%) corresponded to the fixed $N_f = 2$ assumption, whereas 54 (58.7%) had a different visible above-ground story count. These findings confirm that $N_f = 2$ should be interpreted as a simplifying benchmark parameter rather than as a project-specific estimate derived from the dataset.

The subsequent frozen-checkpoint sensitivity analysis was conducted on subsets for which the corresponding manual metadata permitted an intervention. Orientation correction was evaluated for the 13 resolved projects whose manually reviewed facade direction differed from the fixed bottom-edge assumption. Story-count correction was evaluated for the 54 projects with a known visible above-ground story count other than two. A combined orientation-and-story intervention was applicable to 33 projects for which both correction criteria were available. In all cases, only the Skeleton conditioning representation was modified; the trained ControlNet checkpoint and all other inference settings remained unchanged.

Across these post hoc interventions, correcting the fixed Skeleton-construction assumptions did not produce a broad or uniform change across the paired-reference evaluation metrics. The direction and magnitude of metric changes varied across interventions and metrics, indicating that project-specific correction of orientation or story count does not translate straightforwardly into improved similarity to the paired exterior photograph. In the combined intervention, the SSIM comparison yielded an uncorrected p-value of 0.0424; this isolated result is not interpreted as evidence of general superiority, particularly because the analysis was post hoc and involved multiple outcomes. Overall, the sensitivity analysis therefore does not support a claim of systematic metric improvement following metadata correction, but it demonstrates that the frozen Skeleton pipeline can accept revised orientation and story-count parameters without retraining.

These findings clarify the role of the two fixed assumptions in the original benchmark. The audit shows that front_edge = "bottom" and $N_f = 2$ are not universally representative of the test projects, particularly for story count, and they should therefore be regarded as simplifying experimental settings. At the same time, the post hoc interventions show that these parameters can be modified at the conditioning-representation level while retaining the same trained Skeleton ControlNet. This property concerns the editability of the intermediate representation and should not be interpreted as evidence that the generated image will faithfully realize each corrected parameter or that project-specific correction necessarily improves paired-reference image metrics.

4.5. Representational Editability Analysis

The post hoc editability analysis showed that the investigated Skeleton-construction parameters could be modified explicitly at inference time without retraining or adapting the frozen Skeleton ControlNet. Across Projects 245, 463, and 710, changing floor_count from 1 to 2 and 3 produced the intended one-, two-, and three-story organizations in the intermediate Skeleton representations. The front_edge parameter could likewise be reassigned among bottom, left, and right; however, the resulting two-dimensional facade-oriented representation was dependent on the geometry extracted from the individual floor plan and did not necessarily remain distinguishable for every parameter value.

The corresponding generated exterior images did not reproduce the edited parameters with deterministic fidelity. In the floor-count series, substantial changes in the represented number of stories were clearly present in the Skeleton conditions, whereas the generated facades did not consistently reproduce the requested one-, two-, and three-story organizations. The outputs nevertheless varied in facade composition, massing, openings, and surface organization in response to the modified conditioning representations. The facade direction interventions showed a further limitation of the simplified projection. For Project 463, the left- and right-edge settings produced byte-identical Skeleton conditioning images because the extracted side-edge geometry collapsed to the same horizontal feature positions in the two-dimensional front projection. With the same deterministic seed, these identical conditions consequently produced identical generated outputs. This behavior represents a limitation of the current projection scheme rather than a software or caching error.

Figure 3 illustrates the intermediate-representation interventions for Project 463, the predefined 50th-percentile project in the three-project editability sample. The upper series shows the direct modification of the Skeleton representation as the floor-count parameter is varied from one to three. The lower series shows the corresponding facade direction intervention. For this project, the left- and right-edge settings collapse to the same two-dimensional Skeleton representation, illustrating a limitation of the simplified front-projection scheme.

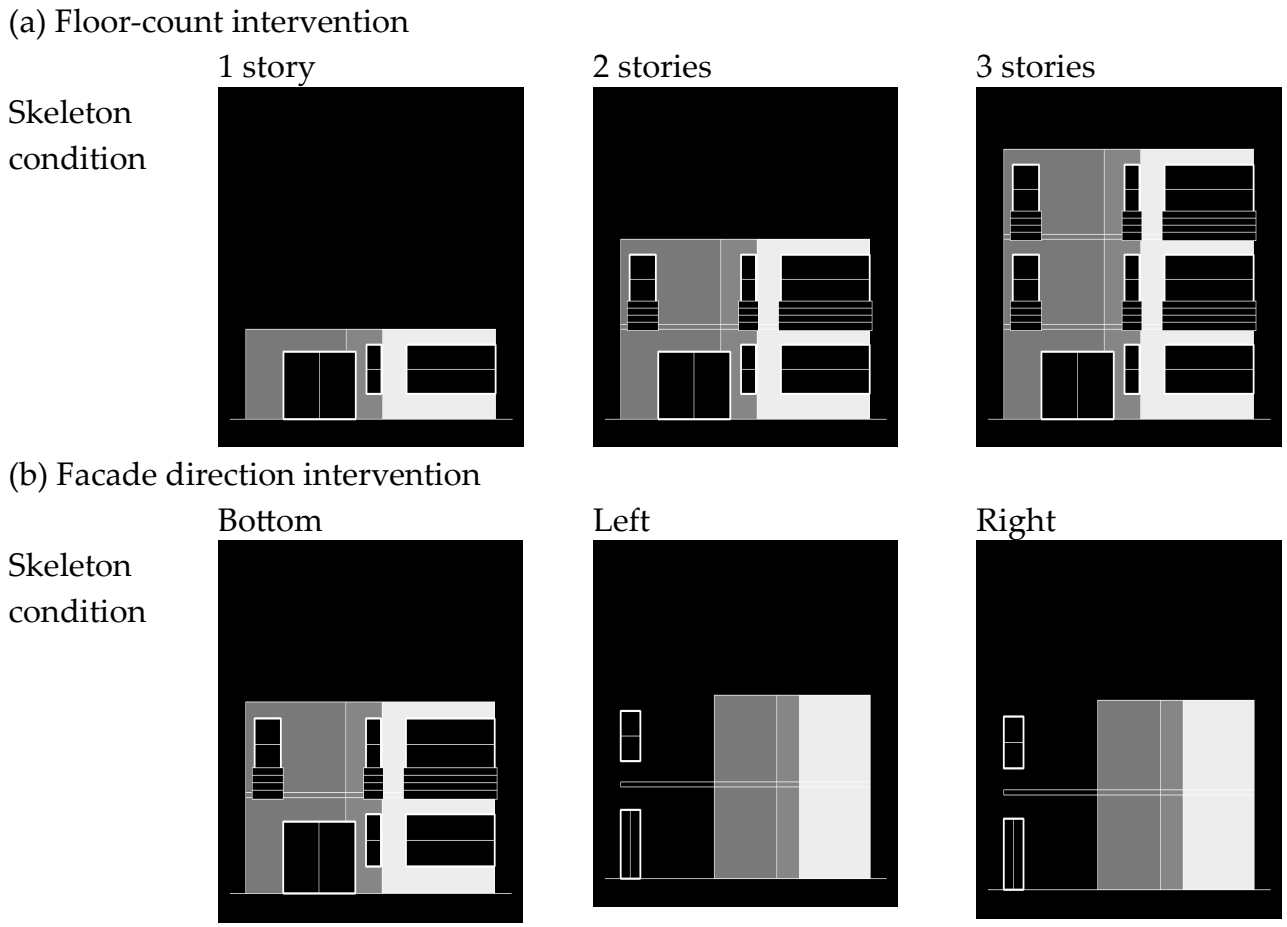


Figure 3. Intermediate-representation editability for Project 463, the predefined 50th-percentile project in the three-project editability sample. (**a**) Floor-count intervention with floor_count varied from 1 to 3 while front_edge = "bottom" is held constant. (**b**) Facade direction intervention with front_edge varied among bottom, left, and right while floor_count = 2 is held constant. All other Skeleton-generation settings remain unchanged. The floor-count intervention produces distinct one-, two-, and three-story Skeleton representations, whereas the left- and right-edge settings collapse to the same two-dimensional representation for this project, illustrating a limitation of the simplified projection scheme.

These results demonstrate representational editability rather than precise output controllability. Unlike the Plan-Edge condition, which is derived directly from the floor plan edge representation, the parametric facade skeleton exposes explicit intermediate parameters that can be modified before generation while retaining the same trained ControlNet. However, the presence of an editable parameter does not imply that its intended architectural effect will be reproduced faithfully by the diffusion model. This observation is consistent with the conditioning adherence results in Section 4.2, which likewise showed incomplete realization of the structural information encoded in the Skeleton condition.

Taken together with the primary benchmark, the editability analysis clarifies the specific and bounded advantage of the proposed Skeleton representation. The Skeleton did not demonstrate statistically superior paired-reference performance relative to the simpler Plan-Edge baseline; instead, its experimentally demonstrated distinction lies in providing an explicit parameterized intermediate representation through which selected structural assumptions can be inspected and modified without retraining the generative model. The results therefore distinguish three separate properties: paired-reference benchmark performance, representational editability, and conditioning adherence. The present findings

support the second of these properties while showing that editability alone does not guarantee the third.

4.6. Deterministic Parser Validation Results

The separate 40-project manual validation showed that the deterministic parser provided only moderate correspondence for the principal exterior perimeter and substantially more limited performance for exterior-opening detection. Across the validation sample, the mean perimeter IoU was 41.05% ± 28.90%, with a median of 41.83%. The mean boundary Chamfer Distance between the automatically extracted and manually annotated exterior perimeters was 51.01 ± 33.65 pixels, with a median of 42.95 pixels. The relatively large dispersion of both measures indicates substantial project-to-project variability in the quality of the extracted perimeter geometry.

Exterior-opening detection was more limited. For exterior windows, micro-averaged precision, recall, and F1-score were 0.2754, 0.1712, and 0.2111, respectively. For exterior doors, the corresponding values were 0.1081, 0.0588, and 0.0762. These results indicate that the deterministic parser should be interpreted as a lightweight heuristic geometry-extraction stage rather than as a high-accuracy architectural parsing model. In particular, missed or incorrectly identified openings can propagate into the structural token representation and subsequently into the parametric facade skeleton. The parser-validation results therefore identify the quality of the initial geometric extraction as an important limitation of the present pipeline and a priority for future methodological improvement.

5. Discussion

5.1. Algorithmic and Architectural Limitations

The system proposed in this study has several algorithmic and architectural limitations. First, it was developed primarily using contemporary residential projects characterized predominantly by rectilinear floor plan geometries. Consequently, its generalizability to curvilinear, organic, free-form, or geometrically irregular architectural configurations has not yet been established. Because cross-sectional information was not available in the source dataset, vertical characteristics that cannot be recovered metrically from floor plans were represented using predefined image space proportions rather than physical dimensions. Accordingly, the "depth-aware" representation used in this study encodes simplified relative depth derived from normalized plan geometry and should not be interpreted as metric three-dimensional reconstruction. The locked benchmark also used the simplifying assumptions front_edge = "bottom" and $N_f = 2$. The post hoc metadata audit showed that these assumptions were not universally representative of the independent test projects, particularly for story count. In addition, the deterministic floor plan parser is a heuristic geometric extraction procedure rather than a learned architectural understanding model, and its separate manual validation showed only moderate perimeter correspondence and limited opening-detection performance. Errors or omissions introduced at this stage can therefore propagate into the structural tokens and subsequent Skeleton representation.

A second limitation concerns the composition and scale of the dataset. Pair-integrity screening of the source file collections yielded 657 complete floor plan–exterior photograph pairs for the final experimental dataset. These were partitioned at the project level into 459 training, 99 validation, and 99 independent test projects. The dataset was derived from contemporary residential projects available through the ArchDaily platform and therefore cannot be assumed to represent the full geographical, cultural, stylistic, or typological diversity of architectural design. Future research should evaluate the framework using larger multi-source datasets encompassing different geographical regions, architectural periods, and building typologies.

The quantitative results also indicate limitations in the performance of the proposed conditioning representation. Under the controlled experimental protocol, the Skeleton ControlNet did not demonstrate statistically significant superiority over the Plan-Edge ControlNet for Chamfer Distance, Edge IoU, SSIM, or CLIP cosine similarity after correction for multiple comparisons. Relative to the text-only Stable Diffusion baseline, Skeleton conditioning produced a statistically significant improvement in Edge IoU, but no significant improvement was observed for Chamfer Distance or SSIM. Moreover, the text-only baseline achieved significantly higher CLIP cosine similarity. These findings indicate that the proposed structural conditioning should not be interpreted as providing uniform improvement across geometric and semantic evaluation criteria. The post hoc editability analysis identifies a more specific distinction from direct Plan-Edge conditioning: the Skeleton provides explicit intermediate parameters that can be inspected and modified without retraining the generative model. However, this representational editability does not establish superior image generation performance or precise output controllability.

The evaluation itself presents additional limitations. The independent test set contains 99 projects, and FID was therefore treated only as a descriptive dataset-level indicator rather than as a basis for inferential statistical comparison. Chamfer Distance, Edge IoU, and SSIM compare generated images with paired project exterior photographs that are not geometrically registered orthographic elevations; these measures should therefore be interpreted as paired-reference similarity proxies rather than as direct measures of floor plan fidelity or architectural correctness. CLIP cosine similarity primarily reflects semantic correspondence between the generated image and the textual prompt and likewise does not measure structural adherence to the conditioning representation. To address this distinction, the present study additionally evaluated story organization and opening-column correspondence directly between the Skeleton conditions and generated outputs. However, this bounded human evaluation considers only selected observable structural attributes and does not constitute a comprehensive measure of architectural, functional, or perceptual quality.

Although the revised evaluation includes a bounded human annotation study of conditioning adherence and an independent second-rater reliability assessment, it does not constitute a perceptual, preference, or usability study involving architects or other design professionals. Consequently, claims regarding architectural preference, design usefulness, creativity, or practical integration into professional workflows cannot be established from the present experiments. The deterministic rule-based component should likewise be understood as an auxiliary, transparent representation of selected author-defined architectural relationships rather than as a learned or empirically validated explanatory model. Its thresholds were not statistically calibrated, and the component does not explain the internal decision-making or latent operations of Stable Diffusion or ControlNet. Future research should therefore incorporate expert evaluation, preference-based and usability studies, broader architectural datasets, and empirical validation of rule-based decision-support relationships.

A further methodological limitation concerns the scope of the baseline comparison. The controlled experiments compare the proposed Skeleton ControlNet primarily with a text-only SD 1.5 baseline and a Plan-Edge ControlNet trained under the same data split, base model, and optimization protocol. This design isolates the effect of the conditioning representation while minimizing confounding differences between model families. However, the study does not include an independently implemented contemporary architectural or layout-conditioned generative model from the literature. Consequently, the reported comparisons should be interpreted as controlled within-framework experiments rather than as

evidence of state-of-the-art performance. Benchmarking against independently developed architectural generation methods constitutes an important direction for future work.

An additional source of uncertainty arises from the physical relationship between the paired floor plans and exterior photographs. The dataset associates a selected floor plan with a representative exterior photograph from the same architectural project; however, this project-level pairing does not establish one-to-one geometric correspondence or a consistently recoverable orientation relationship between the two representations. The manual test set audit reinforced this limitation: a unique facade direction could be resolved for only 56 of the 99 projects, and 13 of those 56 resolved cases were inconsistent with the fixed bottom-edge benchmark assumption. In multi-story buildings, the selected plan may also describe only part of the geometry visible in the complete exterior image, while perspective projection, camera position, occlusion, landscaping, and surrounding context further affect the photographic representation. The paired-reference photographs should therefore be regarded as project-level visual references rather than geometrically equivalent architectural elevations. This uncertainty may contribute to the limited absolute values and cross-metric differences observed for Edge IoU, Chamfer Distance, and SSIM.

The final comparison used matched project-specific deterministic seeds to isolate differences between conditioning strategies; repeated inference runs with multiple seeds per project were not conducted, and seed-to-seed variability was therefore not quantified.

The representational editability analysis also revealed a limitation of the simplified facade-projection procedure. Although explicit Skeleton parameters can be modified without retraining the ControlNet, different parameter settings do not necessarily produce unique conditioning representations for every plan geometry. For Project 463, for example, the left- and right-edge settings collapsed to the same two-dimensional Skeleton representation because the extracted side-edge geometry projected to identical horizontal feature positions. More generally, even when parameter modifications produce distinct Skeleton conditions, the conditioning adherence results show that these changes are not necessarily reproduced faithfully in the generated exterior image. Thus, parameter-level editability should be distinguished from both representational uniqueness and output-level controllability.

5.2. Methodological and Algorithmic Discussion

This study presents an integrated decision-support framework designed to generate floor plan-conditioned exterior building images from two-dimensional floor plans while separately representing selected architectural relationships through an auxiliary deterministic rule-based layer. The final experimental dataset comprised 657 valid floor plan–exterior photograph pairs retained after project-ID-based pair-integrity verification. Project-level partitioning resulted in 459 training, 99 validation, and 99 independent test projects, thereby preventing project overlap between model development and final evaluation.

For conditional facade synthesis, the experiments were conducted using Stable Diffusion 1.5 as the common pretrained generative backbone and Canny-based ControlNet as the initial conditioning architecture. Rather than training separate generative backbones, two ControlNet models were trained under the same experimental protocol using different conditioning representations: direct floor plan-derived edge maps (Plan-Edge ControlNet) and the proposed parametric facade skeleton representation (Skeleton ControlNet). Both models used identical project-level partitions, training duration, base model, and random seed, allowing the effect of the conditioning representation to be examined under controlled conditions. During final inference, identical prompts, inference parameters, and project-specific deterministic seeds were used across the three evaluated configurations: text-only Stable Diffusion 1.5, Plan-Edge ControlNet, and Skeleton ControlNet. This controlled design supports comparison of the conditioning representations within the present framework

but does not by itself establish superiority over independently developed architectural generation methods.

5.2.1. Effect of Structural Conditioning on Edge Correspondence

The quantitative findings provide a more nuanced interpretation of the contribution of structural conditioning. Relative to the text-only baseline, Skeleton ControlNet achieved a higher mean Edge IoU (0.0611 versus 0.0574), and the paired comparison remained statistically significant after Holm correction (adjusted $p = 0.0310$). However, no statistically significant difference was observed for Chamfer Distance. These findings indicate that skeleton conditioning affected paired-reference edge overlap under the present evaluation protocol, but this effect was metric-specific and should not be generalized as overall geometric superiority.

Importantly, the direct comparison between Skeleton ControlNet and Plan-Edge ControlNet did not reveal statistically significant differences in Chamfer Distance, Edge IoU, SSIM, or CLIP cosine similarity after correction for multiple comparisons. Although the Skeleton representation produced slightly higher numerical means for Edge IoU, SSIM, and CLIP cosine similarity, while Plan-Edge conditioning produced a slightly lower Chamfer Distance, these differences were small and statistically non-significant. Accordingly, the present results do not establish the parametric facade skeleton as quantitatively superior to direct Plan-Edge conditioning. Instead, the post hoc editability analysis identifies a more specific representational distinction: the Skeleton exposes explicit intermediate parameters, including represented floor count and facade direction, that can be inspected and modified without retraining the generative model. This distinction is bounded, however, because parameter modifications do not necessarily produce unique conditioning representations for every plan geometry and are not necessarily realized faithfully in the generated exterior image. The conditioning adherence analysis supports this distinction, showing moderate realization of the requested story organization but substantially more limited opening-column correspondence. Thus, the contribution of the Skeleton lies in providing an explicit and editable intermediate conditioning representation rather than in demonstrated metric superiority or guaranteed output-level control. Whether this representational distinction provides additional measurable advantages under different datasets, architectural typologies, conditioning strengths, or model architectures requires further investigation.

5.2.2. Image-Level and Semantic Quality

The image-level results show that the use of explicit structural conditioning does not translate directly into greater similarity across all visual criteria. SSIM remained statistically comparable among the tested conditions, whereas the text-only reference achieved higher CLIP cosine similarity. The descriptive FID values were lower for both conditioned models than for the text-only condition, but the small evaluation set precludes strong interpretation of these differences. Taken together, these results indicate that structural conditioning changes the distribution and edge organization of generated exterior images without demonstrating a uniform improvement in perceptual or semantic image quality.

In the direct comparison between Skeleton ControlNet and Plan-Edge ControlNet, neither SSIM nor CLIP cosine similarity differed significantly after correction for multiple comparisons. The numerical differences between the two conditioned configurations were small, indicating that the additional depth-aware skeleton representation did not produce a measurable advantage in these image-level metrics under the present experimental setting. The significantly higher CLIP cosine similarity of the text-only reference further indicates that the modest improvement in paired-reference Edge IoU observed with Skeleton conditioning was accompanied by lower semantic similarity to the fixed evaluation prompt,

reinforcing that the evaluated metrics capture distinct and potentially non-aligned aspects of the generated images.

5.2.3. Interpretable Rule-Based Decision Support

A further component of the proposed framework is the integration of generative modeling with an auxiliary deterministic rule-based decision-support layer. This component expresses selected predefined architectural relationships through explicit "IF… THEN…" rules derived from structural descriptors. The rules and their numerical thresholds are author-defined heuristics rather than learned, statistically calibrated, or empirically validated architectural relationships. Accordingly, this layer should not be interpreted as a surrogate model or as an explanation of the internal latent decision-making process of Stable Diffusion or ControlNet. Its interpretability is limited to the transparency and traceability of the explicitly encoded rule structures surrounding the generative workflow. Within this bounded scope, the rule-based layer provides a reproducible mechanism for expressing selected architectural relationships as part of the broader decision-support framework. Future research should validate these relationships and thresholds through expert-based and empirical evaluation and investigate whether the rule-based representations provide practical benefits across broader architectural datasets, building typologies, and professional design workflows.

6. Conclusions

This study presents ArchAI 2.0 as an exploratory proof-of-concept decision-support framework that combines deterministic floor plan geometry parsing, structural token extraction, parametric skeleton generation, and conditional diffusion-based exterior image synthesis, together with an auxiliary deterministic rule-based decision-support layer. Following pair-integrity verification, the final experimental dataset comprised 657 valid floor plan–exterior photograph pairs, partitioned at the project level into 459 training, 99 validation, and 99 independent test projects. The controlled evaluation compared text-only Stable Diffusion 1.5, direct Plan-Edge ControlNet conditioning, and the proposed Skeleton ControlNet conditioning under matched inference settings.

The findings indicate that explicit structural conditioning affects different evaluation dimensions in different ways rather than producing uniform improvement across all metrics. Compared with the text-only baseline, Skeleton ControlNet achieved a statistically significant increase in paired-reference Edge IoU after correction for multiple comparisons, indicating greater overlap between edge structures extracted from the generated exterior images and the corresponding reference photographs under the present evaluation protocol. However, no statistically significant improvement was observed for Chamfer Distance or SSIM, while the text-only baseline achieved significantly higher CLIP cosine similarity. Furthermore, the direct comparison between Skeleton and Plan-Edge ControlNet revealed no statistically significant differences across the evaluated sample-level metrics. Therefore, the present results do not establish the parametric facade skeleton as quantitatively superior to direct Plan-Edge conditioning. Instead, the post hoc editability analysis demonstrates a more specific distinction: the Skeleton provides an explicit parameterized intermediate representation in which selected structural assumptions can be inspected and modified without retraining the generative model.

At the same time, the conditioning adherence evaluation demonstrates that representational editability should be distinguished from output-level realization. Story organization was reproduced exactly in 58 of 98 evaluable cases (59.18%), whereas opening-column correspondence was more limited, with a micro-averaged F1-score of 0.2970 at the primary matching threshold of $\tau = 0.08$. The editability analysis similarly showed that explicit

parameter modifications do not necessarily produce unique conditioning representations for every plan geometry or faithful realization of the intended architectural effect in the generated image. Accordingly, the Skeleton should be interpreted as an editable structured conditioning interface rather than as a mechanism for precise or guaranteed architectural control. Separately, the auxiliary deterministic rule-based component expresses selected predefined architectural relationships through human-interpretable "IF... THEN..." rules. These author-defined rules provide a transparent and traceable decision-support representation but do not explain the internal latent decision-making process of Stable Diffusion or ControlNet and have not been empirically validated as universal architectural relationships.

Future research should investigate whether the proposed structural representation provides measurable advantages across larger and more diverse architectural datasets, different building typologies, free-form and non-orthogonal geometries, and alternative conditioning or generative architectures. Further development should improve and validate the deterministic floor plan parsing stage, incorporate richer vertical and sectional information, and investigate project-specific facade direction and story representations beyond the simplifying assumptions used in the present benchmark. The simplified facade-projection procedure should also be refined to reduce cases in which different parameter settings collapse to the same two-dimensional conditioning representation. Finally, expert-based architectural evaluation and user studies are required to assess the practical usefulness, interpretability, and design value of both the structured conditioning representation and the auxiliary rule-based decision-support layer within real architectural workflows.

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

Funding: This study was supported by Istanbul Topkapi University within the scope of its publication incentive program. The funding number is not applicable.

Data Availability Statement: The raw architectural floor plan images and exterior photographs compiled in this study were collected from publicly accessible project pages on ArchDaily using a Playwright web scraper. Because the source images constitute third-party copyrighted material, the raw dataset images are not publicly redistributed. Data-processing and evaluation scripts, model architecture configurations, extracted structural token JSON samples, prompt templates, seed settings, and dataset project URL manifests are available from the corresponding author (ecantemir@gelisim.edu.tr) upon reasonable academic request for non-commercial research purposes.

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

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

References

1. Chatzikonstantinou, I.; Sariyildiz, I.S. Addressing design preferences via auto-associative connectionist models: Application in sustainable architectural Façade design. *Autom. Constr.* **2017**, *83*, 108–120. [CrossRef]
2. Wang, B.; Lu, W.; Zhang, Y. A graph-enabled parametric modeling approach for façade layout generative design. *J. Build. Eng.* **2025**, *105*, 112481. [CrossRef]
3. Newton, D. Generative deep learning in architectural design. *Technol.|Archit. Des.* **2019**, *3*, 176–189. [CrossRef]
4. As, I.; Pal, S.; Basu, P. Artificial intelligence in architecture: Generating conceptual design via deep learning. *Int. J. Archit. Comput.* **2018**, *16*, 306–327. [CrossRef]

5. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 3813–3824.
6. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
7. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
8. LeCun, Y. Generalization and network design strategies. In *Connectionism in Perspective*; North Holland: Haarlem, The Netherlands, 1989; Volume 19, pp. 143–155.
9. Ciresan, D.C.; Meier, U.; Masci, J.; Gambardella, L.M.; Schmidhuber, J. Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the International Joint Conference on Artificial Intelligence*; IJCAI Organization: Barcelona, Spain, 2011; Volume 22, pp. 1237–1242.
10. Kayalı, N.Z.; Omurca, S.İ. Konvolüsyonel sinir ağları ile Çin sayı örüntülerinin sınıflandırılması. *J. Comput. Sci.* **2021**, 184–191. [CrossRef]
11. Mienye, I.D.; Swart, T.G.; Obaido, G. Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information* **2024**, *15*, 517. [CrossRef]
12. Mortezapour Shiri, F.; Perumal, T.; Mustapha, N.; Mohamed, R. A Comprehensive Overview and Comparative Analysis on Deep Learning Models. *J. Artif. Intell.* **2024**, *6*, 301–349. [CrossRef]
13. Yıldız, B.; Aktaş, B. Mimari tasarım sürecinde karar verme: Bulanık mantık tabanlı cephe modeli önerisi. In *11th Computational Design in Architecture National Symposium*; Middle East Technical University: Ankara, Turkey, 2017; Volume 173.
14. Yoshimura, Y.; Cai, B.; Wang, Z.; Ratti, C. Deep learning architect: Classification for architectural design through artificial intelligence. In *International Conference on Computers in Urban Planning and Urban Management*; Springer: Cham, Switzerland, 2019; pp. 249–265.
15. Cantemir, E.; Kandemir, O. Use of artificial neural networks in architecture: Determining the architectural style of a building with a convolutional neural networks. *Neural Comput. Appl.* **2024**, *36*, 6195–6207. [CrossRef]
16. Huang, X.; Zheng, Y. Architectural drawings recognition and generation through machine learning. In *Proceedings of the 38th Annual Conference of the Association for Computer Aided Design in Architecture*; ACADIA: Mexico City, Mexico, 2018; pp. 18–20.
17. Huang, J. Automatic Extraction and Reconstruction of Drawing Design Elements Based on Computer Vision and Neural Networks. *Comput.-Aided Des. Appl.* **2024**, *21*, 50–65.
18. Fan, X.; Wang, Q.; Tang, Q.; He, Y. A Scene Rendering Algorithm for the Fusion of CAD Data and Neural Networks in Environmental Art Design. *Comput.-Aided Des. Appl.* **2024**, *21*, 1–16.
19. Wang, Z.; Wang, D. Deep learning-based automatic optimization of design smart home. *Comput.-Aided Des. Appl.* **2024**, *21*, 96–113.
20. Ruiz-Montiel, M.; Boned, J.; Gavilanes, J.; Jimenez, E.; Mandow, L.; Perez-de-la-Cruz, J.L. Design with shape grammars and reinforcement learning. *Adv. Eng. Inform.* **2012**, *27*, 230–245. [CrossRef]
21. Stump, G.M.; Miller, S.W.; Yukish, M.A.; Simpson, T.W.; Tucker, C. Spatial grammar-based recurrent neural network for design form and behavior optimization. *J. Mech. Des.* **2019**, *141*, 124501. [CrossRef]
22. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27, pp. 2672–2680.
23. Yu, Q.; Malaeb, J.; Ma, W. Architectural facade recognition and generation through generative adversarial networks. In *International Conference on Big Data, Artificial Intelligence and Software Engineering, Bangkok, Thailand, 30 October–1 November 2020*; IEEE: New York, NY, USA, 2020; pp. 310–316.
24. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: New York, NY, USA, 2022; pp. 10674–10685.
25. Ma, H.; Zheng, H. Text semantics to image generation: A method of building facades design base on stable diffusion model. In *Proceedings of the International Conference on Computational Design and Robotic Fabrication, Shanghai, China, 6–7 July 2024*; Springer Nature: Singapore, 2024; pp. 24–34.
26. Kuang, Z.; Zhang, J.; Li, Y.; Fukuda, T. Preserving architectural heritage in urban renewal: A stable diffusion model framework for automated historical facade generation. *npj Herit. Sci.* **2025**, *13*, 256. [CrossRef]
27. Burkart, N.; Huber, M.F. A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.* **2021**, *70*, 245–317. [CrossRef]
28. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]

29. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
30. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*; PmLR: New York, NY, USA, 2021; pp. 8748–8763.

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.