# RETHINKING FLOORPLAN GENERATION WITH SPACE SYNTAX: FROM DESIGN THEORY TO MODEL IMPROVEMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Architectural design increasingly adopts generative AI (GenAI) for floorplan generation, an early-stage design task exploring how spaces should be organized for use. Yet fidelity objectives inherited from native visual and geometric generators do not automatically capture this design purpose. This organization remains largely implicit in generated representations, limiting direct evaluation and feedback. Drawing on space syntax, an established architectural design theory, the Space Syntax Integration Operator (SSIO) recovers convex-space access graphs across heterogeneous formats. This theory-grounded shared intermediate representation, which we call a design container, supports evaluation and model understanding of spatial design. Applying SSIO to four residential corpora spanning 15 countries reveals public spaces' integration levels are highest, with corpus-dependent distributions. These patterns underpin the RPLAN-SSI benchmark, an empirical design reference system assessing population-level functional-profile alignment (IPD) and within-plan public-space advantage (PIA). Comparison with this residential reference reveals a design gap: public spaces in generated floorplans have a smaller integration advantage over other spaces. To address this gap, Space Syntax-Guided Post-Training (SSPT) provides a plug-and-play design-learning interface: SSIO-derived RL feedback improves IPD and PIA across five heterogeneous diffusion backbones while preserving native mechanisms. We further discuss how this approach may inform GenAI applications in other design domains.

## 1 INTRODUCTION

Floorplan generation brings generative AI (GenAI) into architectural design, where the goal is to create spatial layouts that support human use. This task supports early-stage exploration of spatial organization: how spaces support everyday activities, shared use, and privacy (Wu et al., 2019; Park et al., 2024). Many existing methods inherit the training objectives of their underlying GenAI models, improving visual and geometric quality; architectural requirements typically enter as input conditions (Hu et al., 2020; Nauata et al., 2021; Shabani et al., 2023). Outputs use diverse visual and geometric formats, while the spatial organization that matters for use remains largely implicit. Common evaluation metrics assess visual similarity, geometric validity, and constraint satisfaction (Regenwetter et al., 2023; Zeng et al., 2025a), but do not directly measure how well spaces are organized for use (Figure 1(a)). For example, a connected generated floorplan can give a bathroom higher overall accessibility than its shared living space (Figure 7, B). Quantitative design theory and residential evidence accumulated in architecture let us test design-directed improvements across generators while retaining their original visual fidelity objectives.

Recent work begins to assess design performance: FloorPlan-LLaMa learns architectural judgments from expert preferences (Yin et al., 2025). Its evaluator, however, predicts design scores without explicitly representing the spatial objects and relations to be measured. Architectural design theories offer a way to make these objects and relations explicit and explain why they matter for use. Recovering these objects and relations across formats would make spatial organization directly measurable, providing a common architectural basis for identifying design gaps, setting improvement objectives, and guiding learning across heterogeneous generators.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

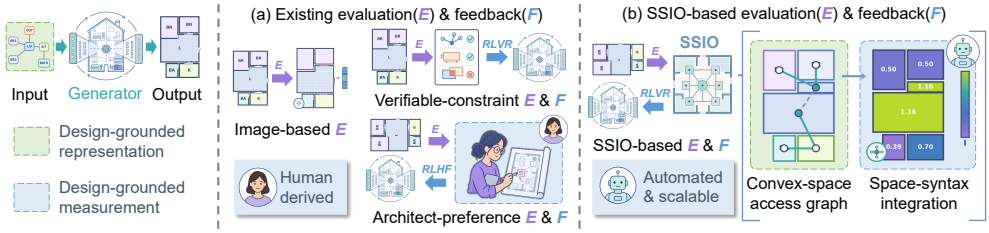



Figure 1: **Towards Design-directed Representation and Measurement in Floorplan Generation.**

For this purpose, we draw on space syntax, an established architectural design theory linking spatial organization to building use: each space's role depends on its relations within the overall layout (Hillier & Hanson, 1984; Hillier, 1996; Hanson, 1998). It represents interiors as convex-space access graphs (CSAGs): convex spaces are nodes, access relations are edges, and integration measures relative accessibility. Our Space Syntax Integration Operator (SSIO) recovers CSAGs from heterogeneous sources and computes integration scores (Figures 1(b) and 2). This yields a shared intermediate representation (IR) of design elements and relations—which we term a **theory-defined design container**—and directly computable spatial measurements across generators.

To ground these measurements empirically, we apply SSIO to four residential corpora spanning 15 countries (Wu et al., 2019; LIFULL Co., Ltd., 2015; van Engelenburg et al., 2025; Ziaee et al., 2024). Public spaces have the highest integration levels in each corpus (Figure 3). This shared Public-first ordering grounds Public anchoring, while differing distributions call for references matched to the residential context and task. Accordingly, the RPLAN-derived Space Syntax Integration benchmark (RPLAN-SSI) uses RPLAN's Chinese residences as its reference: Integration Profile Distance (IPD) measures population-level functional-profile alignment, and Public-Anchor Integration Advantage (PIA) measures within-plan Public anchoring. A room-count-matched comparison of the output population from a generator pretrained on RPLAN against this corpus reference reveals preserved Public-first ordering but a substantially smaller Public integration margin over other functional domains (Figure 4). These analyses establish an **empirical design reference system** for benchmarking generators and identifying design gaps.

To close this design gap, we introduce Space Syntax-Guided Post-Training (SSPT), using grounded measurements to improve heterogeneous generators. For diffusion floorplan generators (Shabani et al., 2023; Wang & Pajarola, 2025; 2026), this requires connecting measurements of completed layouts to the native sampling decisions that produce them. Through backbone-specific policy bindings, SSPT applies a shared proximal policy optimization (PPO) objective (Schulman et al., 2017) with SSIO-derived terminal rewards to native reverse policies over selected lower-noise transitions (Figure 5). Across five diffusion backbones, both measures improve; primary results show 27.0% lower IPD and 41.9% higher median PIA within native-quality budgets. A connected case illustrates the gains: shared living-space anchoring replaces bathroom dominance under an unchanged room program (Figure 7). Shared feedback and native-policy bindings thus form a **reusable design-learning interface** for plug-and-play post-training without replacing native mechanisms.

To situate this approach within GenAI for design, we review 55 studies across 14 design categories through three maps comparing design representations and criteria, where feedback acts, and cross-generator reuse. The maps highlight a distinctive connection: theory-defined measurements and empirical references guide heterogeneous-generator post-training through a shared design container. We distill this connection into a three-element framework and executable harness as a reference for design-theory translation in other domains (Section 5.6; Appendices F and G).

Our contributions translate architectural design theory into a representation–evaluation–learning pathway for floorplan generation: **(1)** SSIO makes spatial organization explicit and directly measurable across heterogeneous formats, providing a unified design container for evaluation and feedback. **(2)** RPLAN-SSI establishes an empirical spatial-organization benchmark, revealing weaker Public anchoring in generated layouts than in real floorplans. **(3)** SSPT introduces a reusable design-learning interface—to our knowledge, the first plug-and-play RL post-training framework for diffusion floorplan generators—improving both metrics across five backbones without replacing native mechanisms.

## 2 Related Work

**Floorplan generation.** Floorplan generators combine architectural conditioning with synthesis in different output spaces. Graph2Plan and House-GAN++ realize requested layouts through box prediction and mask generation, whereas HouseDiffusion denoises room polygons (Hu et al., 2020; Nauata et al., 2021; Shabani et al., 2023). Boundary-focused and language-conditioned diffusion vary geometric representations and user controls (Wang & Pajarola, 2025; Hu et al., 2025a; Qin et al., 2024). Requested relations do not determine the access hierarchy realized after synthesis. SSIO measures this realized organization across native output formats.

**Architectural measurement and evaluation.** Distributional similarity, geometric validity, and condition satisfaction assess fidelity and compliance (Shabani et al., 2023; Zeng et al., 2025a; Park et al., 2024). Space syntax supplies analytical objects and accessibility measures for studying how spatial configuration supports use (Hillier & Hanson, 1984; Hillier, 1996; Hanson, 1998). Its applications include design transformation and analysis of visibility or connectivity in generated layouts (Eloy & Guerreiro, 2016; Heo & Cho, 2023; Man et al., 2026). Our challenge is to recover a consistent analytical structure across formats and establish references for interpreting its measurements. SSIO and RPLAN-SSI connect this recovery to residential evidence: theory specifies the property, while real designs ground numerical comparisons of the resulting spatial organization.

**Feedback and post-training.** FloorPlan-LLaMa learns architect-preference rewards, while verifiable-reward RL checks connectivity and area; both update autoregressive generators (Yin et al., 2025; Lara et al., 2026). These approaches bring design judgment or specified requirements into learning. GreenPlanner targets energy and feasibility in a flow generator (Zeng et al., 2026). SSPT instead targets access organization with directly computed architectural measurements. Diffusion alignment provides differentiable-reward optimization and policy gradients (Clark et al., 2024; Black et al., 2024); we adopt policy gradients because geometric recovery and access-graph analysis do not supply reward gradients. Backbone-specific likelihood bindings let the shared feedback update native diffusion policies across different output representations and generation mechanisms.

Our broader comparison examines how design criteria are represented, where they affect learning, and how they transfer across generators (Appendix F). Domain-purpose optimization and multi-backbone adaptation have precedents (Zheng et al., 2023; Lai et al., 2026; Wu et al., 2024b); this study connects theory-defined analysis, empirical references, and heterogeneous-generator post-training. Section 5.6 develops this combination into a reusable theory-to-learning construction workflow.

## 3 Design Representation & Evaluation: Space Syntax System

Given an exterior boundary $B$, room types and counts $\mathcal{M}_c$, and optional intended relations $G_c$, generator $m$ produces a native floorplan $x_m^{\text{nat}}$ conditioned on $c = (B, \mathcal{M}_c, G_c)$. Outputs include raster masks, room boxes, polygons, or structural boundary graphs. To improve functional organization at required quality, SSIO recovers and measures shared structure, RPLAN-SSI compares scores with real designs, and SSPT updates generators (Section 4).

### 3.1 Space Syntax Integration Operator (SSIO)

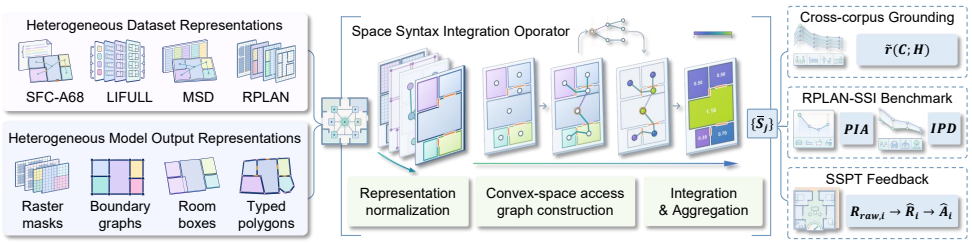



Figure 2: **Space Syntax Integration Operator (SSIO).** Recovering shared theory-defined IR.

Before graph construction, SSIO recovers walkable room regions, their semantic labels, and interior-door incidence from each generator's native output.

Suppressing the plan index, deterministic normalization gives $z_m = \mathcal{A}_m(x_m^{\mathrm{nat}}; c) = \big((Q_{m,j}, y_{m,j})_{j=1}^{R_m}, D_m^{\mathrm{int}}\big)$, where $Q_{m,j}$ is a walkable room region, $y_{m,j}$ its label, and $D_m^{\mathrm{int}}$ the interior-door incidence. A frozen completion map supplies missing door incidence: $\widehat{D}_m^{\mathrm{int}} = \mathcal{C}_\phi(x_m^{\mathrm{nat}}, c)$. SSIO decomposes regions into convex-atom sets $V_j$, connects atoms through within-room continuity and interior doors, retains $\mathcal{R}_m^+ = \{j : |V_j| > 0\}$, and constructs $G_x = (V_x, E_x)$, with $V_x = \biguplus_{j \in \mathcal{R}_m^+} V_j$. The condition graph $G_c$ encodes intended relations, whereas $G_x$ captures realized access.

Node centralities $S_v$ are finite and nonnegative. The benchmark and SSPT use component-local Legacy Hillier–Hanson (HH) integration $S_v = I_v$, with finite boundary conventions (Appendix Eq. (A.2)) (Hillier & Hanson, 1984; Hillier, 1996); larger $I_v$ indicates greater topological integration on the nondegenerate branch. SSIO averages node scores within rooms to obtain $\bar{S}_j$, then room scores within category $k$ of plan $p$ to obtain $\bar{S}_{p,k}$. Under Legacy HH, these are $\bar{I}_j$ and $\bar{I}_{p,k}$, respectively.

### 3.2 RPLAN-derived Space Syntax Integration Benchmark (RPLAN-SSI)

We aggregate SSIO category scores into four functional domains $\mathcal{G}$—Public, Private, Service, and Outdoor—and normalize within plans (Appendix B).

For measurable plan $p$, $\mathcal{H}_p \subseteq \mathcal{G}$ lists its nonempty domains and $\mathcal{K}_{p,g}$ its retained categories in domain $g$. The domain score $S_{p,g}$ averages $\bar{S}_{p,k}$ over $k \in \mathcal{K}_{p,g}$, weighted by retained room counts. With $\epsilon_{\mathrm{norm}} = 10^{-6}$, we define relative domain scores and the functional profile of cohort $\mathcal{C}$, requiring nonempty support in each domain to compute its cohort median:

$$r_{p,g}^{\mathrm{dom}}(S) = \frac{S_{p,g}}{\epsilon_{\mathrm{norm}} + \mathrm{mean}_{h \in \mathcal{H}_p} S_{p,h}}, \qquad \tilde{\mathbf{r}}(\mathcal{C}; S) = \left( \underset{p \in \mathcal{C}:\, g \in \mathcal{H}_p}{\mathrm{median}} \; r_{p,g}^{\mathrm{dom}}(S) \right)_{g \in \mathcal{G}}. \tag{1}$$

The profile collects four domain-wise medians across cohort $\mathcal{C}$. For plans containing Public and another domain, the Public-domain margin $\Delta_{\mathrm{pub}}(p) = r_{p,\mathrm{Public}}^{\mathrm{dom}}(S) - \max_{g \in \mathcal{H}_p \setminus \{\mathrm{Public}\}} r_{p,g}^{\mathrm{dom}}(S)$ measures Public's lead over its strongest competing domain; it is positive exactly when Public ranks strictly first. We suppress the declared scoring rule $S$ in $\Delta_{\mathrm{pub}}(p)$.

**Cross-Corpus Grounding of Functional Regularities.** Which organizational patterns recur across contexts? We apply SSIO to four residential corpora spanning 15 countries: RPLAN (China), LIFULL (Japan), MSD (Switzerland), and SFC-A68 (13 countries). Each corpus retains floorplans with an annotated Public space. Global harmonic centrality, $S_v = H_v$, measures whole-graph accessibility; unreachable spaces contribute zero (Appendix Eq. (A.3)). The functional profiles place Public highest across all four corpora, while Private, Service, and Outdoor remain near or below the within-plan mean (Figure 3). The shared ordering identifies a common design direction; corpus-dependent distributions require task-matched numerical references. Theory determines what to measure, while real designs ground the organization to be learned and the reference for evaluating it.

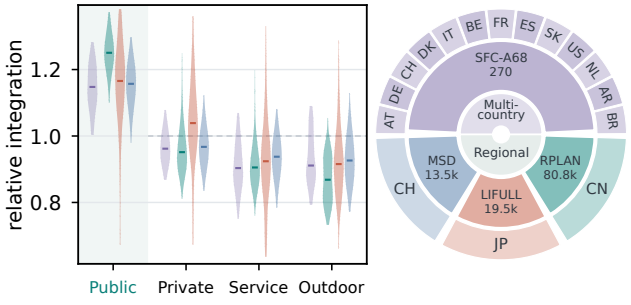



Figure 3: **Cross-Corpus Grounding: Public-first.**

**Configurational Metrics and Evaluation.** RPLAN-SSI uses Legacy HH and selects Living, RPLAN's predominant Public category, as the plan-level anchor. The reference $\mathcal{C}_{\mathrm{ref}}$ contains 20,348 retained exact-eight RPLAN floorplans, matching generated room counts. Each model makes one attempt per condition. Measurable outputs enter the accepted cohort $\mathcal{C}$ if they retain eight spatial instances, including Living, so generators are compared under the same structural requirement.

Integration Profile Distance (IPD) measures the mean absolute mismatch between the accepted and reference functional profiles (Eq. (1)). Public-Anchor Integration Advantage (PIA) complements this population measure with Living's normalized lead over its strongest eligible category competitor.

For plan $p$, let $\mathcal{K}_p$ contain its retained categories, $\mathcal{K}_{\text{norm}}$ the normalization categories, and $\mathcal{V}$ the eligible non-Living competitors. With $\mu_p = \epsilon_{\text{norm}} + \text{mean}_{k \in \mathcal{K}_{\text{norm}} \cap \mathcal{K}_p} \bar{I}_{p,k}$, we define

$$\text{IPD}(\mathcal{C}, \mathcal{C}_{\text{ref}}) = \frac{1}{|\mathcal{G}|} \left\| \tilde{\mathbf{r}}(\mathcal{C}; I) - \tilde{\mathbf{r}}(\mathcal{C}_{\text{ref}}; I) \right\|_1, \qquad \text{PIA}(p) = \frac{\bar{I}_{p,\text{Living}} - \max_{k \in \mathcal{V} \cap \mathcal{K}_p} \bar{I}_{p,k}}{\mu_p}. \quad (2)$$

Lower IPD means closer population alignment. PIA requires an eligible competitor and is positive when Living outranks every eligible category mean; the Public-domain margin instead compares functional domains. We also report acceptance over all attempts. Appendix B.3 defines category membership and the support required to compute each measurement.

**Design Gap between Generated and Real Floorplans.** Under matched room counts and acceptance rules, Legacy-HH functional profiles of HouseDiffusion (HD) outputs and the RPLAN reference reveal preserved Public-first ordering but a weaker hierarchy (Figure 4(a)). HD's median Public-domain margin is 46.3% lower than RPLAN's (0.3387 versus 0.6304); even its 95th percentile remains below the RPLAN median (Figure 4(b)). Yet $\Pr(\Delta_{\text{pub}} > 0)$ is only 4.37 percentage points lower. The gap concerns not just whether Public leads, but by how much, motivating post-training that explicitly improves functional organization.

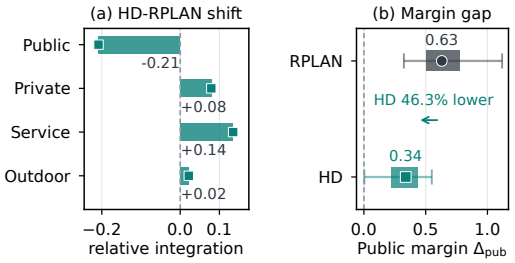



Figure 4: **Preserved Public-first Ordering, Weaker Functional Hierarchy.**

## 4 Design Learning: Space Syntax-Guided Post-Training (SSPT)

SSPT-PPO turns SSIO measurements into terminal feedback, sharing architectural rewards and a PPO objective across native reverse policies. Learning through policy likelihoods lets discrete spatial analysis guide generator updates without differentiating through geometric recovery (Figure 5).

**From Selection to Signed Feedback.** SSPT-Iter post-trains in data space by retraining on SSIO-ranked real and generated examples (Dong et al., 2023; Ferbach et al., 2024). The pilot retains none of 30,000 generated candidates, incurring sampling and scoring costs without learning from generated designs (Appendix E.4.1). SSPT-PPO iterates in trajectory space, using signed advantages to reinforce or suppress decisions according to relative terminal rewards.

**Guarded Terminal Feedback.** We translate the empirical Public-anchor relation into a room-level reward; IPD and PIA assess population- and plan-level outcomes. Fix backbone $m$ and its training measurement settings, suppressing $m$ in reward terms. For a reward-valid output $v_m(x) = 1$, let $\mathcal{R}_{\text{Living}}(x)$ and $\mathcal{R}_{\neg\text{Living}}(x)$ contain retained Living and non-Living rooms. Define $R_{\text{anchor}}(x) = \max_{j \in \mathcal{R}_{\text{Living}}(x)} \bar{I}_j - \max_{j \in \mathcal{R}_{\neg\text{Living}}(x)} \bar{I}_j$, taking empty maxima as zero. A one-sided area guard controls the expansion shortcut exposed by anchor-only optimization: $R_{\text{AreaGuard}}(x) = -[\text{Area}(x) - q_A]_+ / s_A$, where $[u]_+ = \max(u, 0)$, $\text{Area}(x)$ is total retained-atom area, and frozen Base outputs fix threshold $q_A$ and scale $s_A > 0$. The guard weight $\lambda \geq 0$ sets the trade-off between anchoring and expansion control:

$$R^{(\lambda)}_{\text{raw}}(x) = 8R_{\text{anchor}}(x) + \lambda R_{\text{AreaGuard}}(x), \qquad v_m(x) = 1. \quad (3)$$

For $v_m(x) = 0$, set $R^{(\lambda)}_{\text{raw}}(x) = -1$. Appendix C.2 specifies measurement settings, validity, and frozen guard statistics. Below, $R_{\text{raw}} = R^{(\lambda)}_{\text{raw}}$ uses the chosen weight.

**Native Reverse Policies and Rollouts.** Each backbone's selected lower-noise reverse segment defines a finite-horizon Markov decision process with decisions $\mathcal{D}^{(m)}_i = \{1, \ldots, D_m\}$ at decreasing native times $t^{(m)}_1 > \cdots > t^{(m)}_{D_m+1}$. State $s^{(m)}_{i,d}$ contains $x^{(m)}_{i,t^{(m)}_d}$, its time index, condition $c_i$, and the fixed masks or embeddings that define its native generation context.

Joint action $a^{(m)}_{i,d}$ comprises the sampled active stochastic coordinates of $x^{(m)}_{i,t^{(m)}_{d+1}}$, with fixed coordinates restored before the next step. Outside the stochastic decision set, the rollout checkpoint completes generation deterministically to obtain native output $x^{\text{nat}}_{i,m}$.

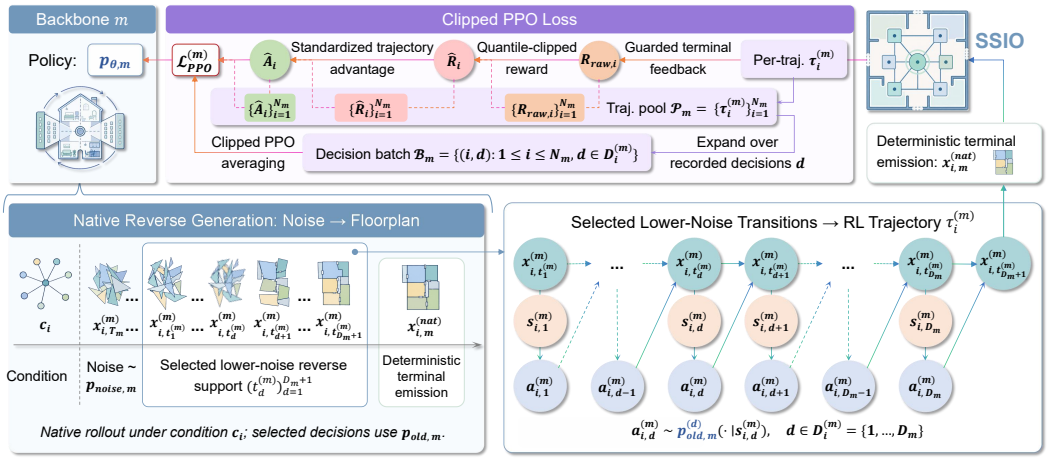



Figure 5: **SSPT-PPO trajectory-space post-training.** Guarded SSIO rewards are clipped and standardized over rollout pools; the resulting trajectory advantages drive PPO updates on backbone-native lower-noise transitions.

At iteration $r$, $p_{\text{old},m} = p_{\theta_r,m}$ samples a fixed trajectory pool $\mathcal{P}_m = \{\tau_i^{(m)}\}_{i=1}^{N_m}$. Terminal outputs receive $R_{\text{raw},i} = R_{\text{raw}}^{(\lambda)}(x_{i,m}^{\text{nat}})$ from Eq. (3); PPO holds outputs and rewards fixed during replay and draws a fresh trajectory pool from the updated policy at the next iteration.

**Clipped Trajectory Optimization.** Let $\mathcal{B}_m = \{(i,d) : 1 \le i \le N_m,\ d \in \mathcal{D}_i^{(m)}\}$ be the decision batch. Rollouts store $\ell_{\text{old},i,d}^{(m)} = \log p_{\text{old},m}^{(d)}(a_{i,d}^{(m)} \mid s_{i,d}^{(m)})$; PPO recomputes $\ell_{\theta,i,d}^{(m)} = \log p_{\theta,m}^{(d)}(a_{i,d}^{(m)} \mid s_{i,d}^{(m)})$ over each binding's active stochastic coordinates. For backbone $m$, log-ratio bound $M_m$ stabilizes the likelihood ratio between the current and rollout policies:

$$\widetilde{\rho}_{i,d}^{(m)} = \exp\Big(\text{clip}\Big[\ell_{\theta,i,d}^{(m)} - \ell_{\text{old},i,d}^{(m)}, -M_m, M_m\Big]\Big). \tag{4}$$

Clipping terminal rewards at pool quantiles $q_{0.01}, q_{0.99}$ gives $\widehat{R}_i = \text{clip}(R_{\text{raw},i}, q_{0.01}, q_{0.99})$. The detached trajectory advantage is $\widehat{A}_i = (\widehat{R}_i - \mu_{\widehat{R}})/(\sigma_{\widehat{R}} + 10^{-8})$, using pool mean $\mu_{\widehat{R}}$ and population standard deviation $\sigma_{\widehat{R}}$. For PPO clip radius $\epsilon$, the shared surrogate loss is

$$\mathcal{L}_{\text{PPO}}^{(m)} = -\mathbb{E}_{(i,d)\sim\text{Unif}(\mathcal{B}_m)}\Big[\min\Big(\widetilde{\rho}_{i,d}^{(m)}\widehat{A}_i, \text{clip}(\widetilde{\rho}_{i,d}^{(m)}, 1-\epsilon, 1+\epsilon)\widehat{A}_i\Big)\Big]. \tag{5}$$

All recorded decisions in a trajectory share $\widehat{A}_i$; reward quantiles and moments remain fixed during replay. This reinforcement learning with verifiable rewards (RLVR) uses computable architectural measurements to ground model improvement in design theory. Appendix C.2 specifies native policy bindings, reverse-time supports, and training settings.

**Base-Referenced Quality and Model Selection.** We seek design improvement within the native-quality range supported by the unchanged generator. Retrospective calibration uses repeated Base HD measurements alone, so candidate design gains do not set the admissible range. Upper prediction limits account for variation in future five-repeat means, with a joint 5% tail-error budget across three batches and two metrics. This budget controls false exclusion of unchanged Base under the repeat-score model. The resulting Fréchet Inception Distance (FID)/Compatibility ceilings are 9.5/2.8. On held-out validation conditions, we rank checkpoints meeting these limits and acceptance/support requirements by lower IPD, then higher median PIA (Appendix C.3).

## 5 Experiments

**Experimental Protocol.** Eight generators share the frozen RPLAN-SSI benchmark, reporting IPD, median PIA, and Exact-8 acceptance (Appendix D.1).

HD provides controlled quality, ablation, and spatial-interpretation evidence; paired Base/post-training comparisons test five-backbone portability, including the separately analyzed GSDiff-SSIO system. Post-training uses at-most-seven-room programs; final evaluation uses five disjoint blocks of 1,000 exact-eight conditions, independently of checkpoint selection. This tests generalization beyond the room counts used for post-training. Appendices B and C.2 detail the protocols.

### 5.1 Unified Benchmark Evaluation

ChatHouseDiffusion (ChatHD) leads profile alignment and Public anchoring among eight generators (Table 1). HD combines broad coverage (94.5% acceptance) with a clear gap to real designs, supporting further design improvement of an already reliable generator. Figure 6(a) includes GSDiff-SSIO, analyzed in Section 5.5; Appendix E.2 compares results across five independent evaluation blocks.

Table 1: **RPLAN-SSI evaluation.**

| Method | IPD ↓ | median PIA ↑ | Acc. (%) |
|---|---|---|---|
| CE2EPlan | 0.1914 | 0.0726 | 66.5 |
| DiffPlanner | 0.1803 | 0.1113 | 59.1 |
| House-GAN++ | 0.1612 | 0.1583 | 40.7 |
| House-GAN | 0.1377 | 0.2606 | 36.1 |
| Graph2Plan | 0.1078 | 0.2886 | 71.6 |
| FloorplanGAN | 0.0943 | 0.2296 | 91.5 |
| ChatHD | **0.0866** | **0.3295** | 31.2 |
| HD | 0.1111 | 0.2800 | **94.5** |

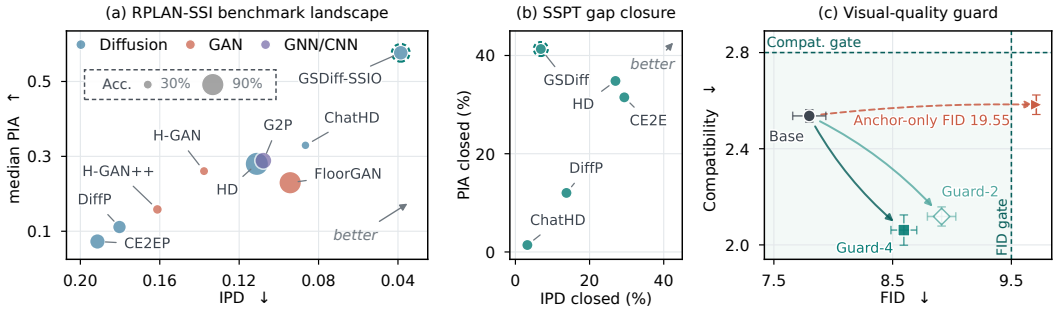



Figure 6: **Design evaluation and post-training.** (a) Model families and Exact-8 acceptance (bubble area). (b) Gap closure to zero IPD and RPLAN median PIA. (c) AreaGuard trade-off: independent Base evaluation, Base-calibrated limits (dashes), and standard deviations (error bars).

### 5.2 Design Improvement under Native-Quality Constraints

**Quality-constrained gains.** Validation selects AreaGuard-4 at iteration 20 for the lowest eligible IPD under the Base-calibrated limits (Appendix E.4.6). Independent testing shows 27.0% lower IPD and 41.9% higher median PIA; acceptance rises from 94.5% to 96.5% (Table 21). FID and Compatibility rise, but remain within the Base-referenced quality limits (Table 2).

Table 2: **Native-quality trade-off of the selected HD checkpoint.** Five-repeat mean ± std.

| Variant | FID ↓ | Compatibility ↓ |
|---|---|---|
| Base HD | $7.798 \pm 0.140$ | $2.537 \pm 0.026$ |
| +SSPT-PPO | $9.087 \pm 0.098$ | $2.572 \pm 0.049$ |
| Upper limit | 9.500 | 2.800 |

**Gains on identical inputs.** Among 4,588 conditions accepted by both models, post-training raises PIA in 70.42% of pairs (median change $+0.1164$; Figure 7). Within-block bootstrap resampling gives a 95% interval of $[0.1082, 0.1257]$ for this median shift. Resampling all attempted conditions gives a 95% interval of 25.2–28.6% for the IPD reduction. Gains persist for connected pairs and alternative centralities, supporting improvement beyond connectivity repair. Public, Private, and Service move toward RPLAN's profile, while Outdoor moves away (Appendices E.3 and E.1). Together, these tests support design improvement at both population and individual-plan levels.

**Spatial interpretation.** Case B replaces bathroom dominance with Living's Public anchoring in connected layouts with the same eight-room program and total area. Case C repairs disconnection and improves anchoring. Cases A, B, and I improve despite fewer cross-room links; Case J adds a link but loses anchoring. Connectivity and link count alone cannot explain these gains.

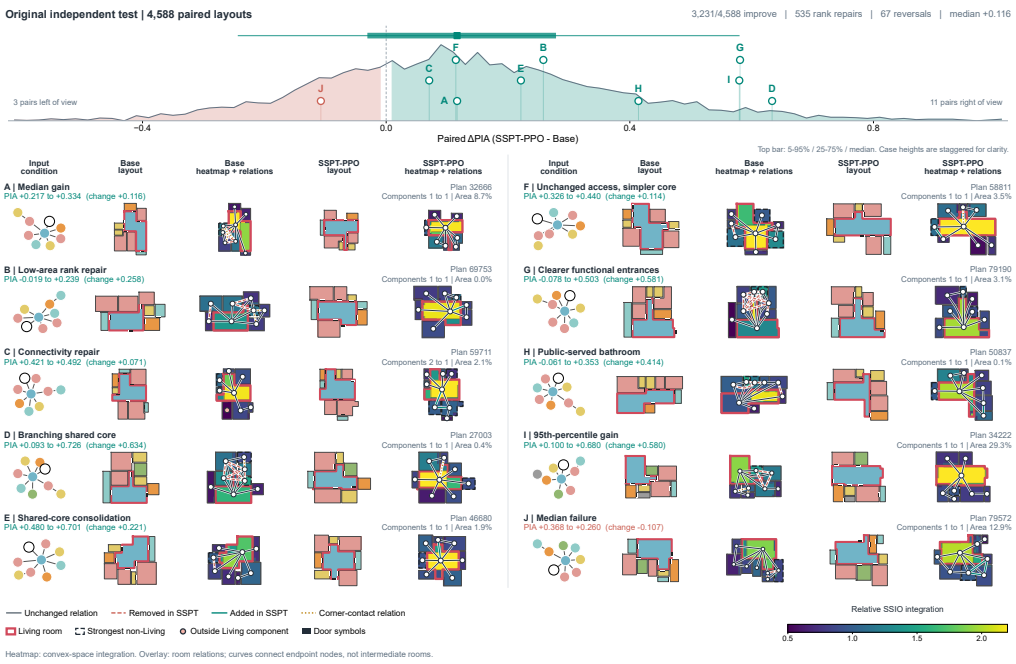



Figure 7: **HD+SSPT-PPO reorganizes functional configuration under fixed programs.** Top: PIA changes for 4,588 matched pairs. Bottom: ten examples; heatmaps encode relative convex-space integration and links indicate room-level access.

### 5.3 Ablation Studies

Controlled HD ablations test update location, design reward, and expansion control (Appendix E.4).

**(1) Where should post-training act?** With profile-distance reward and retained-step budget fixed, lower-noise updates outperform uniform full-horizon and higher-noise updates on both metrics (Table 3). Among these schedules, feedback is most effective late in HD's denoising process.

Table 3: **HouseDiffusion reverse-time ablation.**

| Reverse-time support | IPD ↓ | median PIA ↑ |
|---|---|---|
| Lower-noise segment | **0.0945** | **0.3349** |
| Uniform full horizon | 0.1026 | 0.2996 |
| Higher-noise segment | 0.1126 | 0.2781 |

**(2) Which design reward should guide learning?** Anchor rewards Living's lead over other rooms; category alignment matches category profiles to RPLAN. Living/MAD standardizes Living's score against non-Living category scores using their median and median absolute deviation. The first two perform similarly and outperform Living/MAD; combining them yields the best design scores but exceeds the FID limit (Table 19), motivating expansion control.

**(3) Does controlling expansion preserve design gains?** With Anchor and iteration 10 fixed, AreaGuard-2 retains design gains while sharply reducing FID and Compatibility (Table 4, Figure 6(c)). Weight 4 further improves native quality at some design cost. Area control curbs a reward shortcut: excessive expansion can raise raw anchoring scores while degrading native geometric quality. Guard strength and training duration are then selected jointly.

Table 4: **AreaGuard improves the quality–design trade-off.** Iteration-10 development.

| Variant | IPD ↓ | median PIA ↑ | FID ↓ | Compatibility ↓ |
|---|---|---|---|---|
| Anchor | 0.0830 | 0.3912 | $19.55 \pm 0.18$ | $2.583 \pm 0.041$ |
| +AreaGuard-2 | **0.0827** | **0.3921** | $8.91 \pm 0.12$ | $2.118 \pm 0.040$ |
| +AreaGuard-4 | 0.0858 | 0.3817 | **8.59±0.11** | **2.062±0.063** |

### 5.4 Cross-Backbone Portability

Across five diffusion backbones, SSPT improves population-profile alignment and within-plan Public anchoring (Table 5, Figure 6(b)). Each model is post-trained separately using the same architectural objective. Its policy binding exposes native stochastic decisions and their likelihoods to PPO, accommodating different conditions, representations, and coordinates. Reusing the objective therefore does not require a common output format: theory-defined measurements and residential evidence specify the target, while native policies provide the update mechanisms.

Table 5: **SSPT-PPO across five diffusion backbones.**

| | IPD ↓ | | median PIA ↑ | |
|---|---|---|---|---|
| Backbone | Base | +SSPT-PPO | Base | +SSPT-PPO |
| HD | 0.1111 | **0.0811** | 0.2800 | **0.3974** |
| CE2EPlan | 0.1914 | **0.1352** | 0.0726 | **0.2441** |
| DiffPlanner | 0.1803 | **0.1555** | 0.1113 | **0.1721** |
| ChatHD | 0.0866 | **0.0838** | 0.3295 | **0.3336** |
| GSDiff-SSIO | 0.0386 | **0.0360** | 0.5766 | **0.5934** |

Each Base/post-trained pair uses the same 5,000 conditions and a fixed measurement binding. Acceptance increases for HD, DiffPlanner, and GSDiff-SSIO; CE2EPlan shows substantial metric gains but lower acceptance, while ChatHD changes modestly in metrics and acceptance (Appendix E.5). Design gains thus accompany backbone-dependent coverage, with controlled native-quality evidence on HD. GSDiff additionally tests measurement when generated boundaries leave access unspecified.

### 5.5 Completing Design Relations: The GSDiff Case

GSDiff (Hu et al., 2025a) encodes room boundaries but leaves access undetermined: adjacent rooms need not be directly traversable. A predictor trained on local access labels infers access pairs from geometry, room types, and local context. It is frozen for both Base and +SSPT-PPO; SSIO computes integration on the completed graph. Post-training then improves both design metrics by updating the generator under this fixed completion map (Table 5).

Held-out completion yields access-pair F1 of 0.8879 but median absolute PIA error of 0.3454: accurate local recovery does not ensure accurate whole-plan measurements. On 538 conditions accepted under all completion routes, removing room-type cues lowers Base median PIA from 0.5882 to 0.2781 with geometry fixed. GSDiff-SSIO scores reflect both generated geometry and function-informed access completion (Appendix E.6). The theory identifies information that geometric synthesis can leave unresolved: design feedback requires not only room shapes, but the access relations through which spaces serve their functions.

### 5.6 Design-Theory Translation across Domains

We situate our floorplan approach within a review of 55 studies across 14 design categories (Appendix F). Three comparative maps distinguish what feedback measures, where it acts, and its reuse across generators (Figures 18 and 19). Theory-defined analysis rarely guides post-training across heterogeneous formats in these studies; our floorplan experiments demonstrate this connection. To explore this connection in other design domains, we organize our method's three components—the *design container*, *empirical reference system*, and *design-learning interface*—into a reusable theory-translation workflow and agent-assisted harness (Appendix G). User-interface (UI), scientific-diagram, and poster cases demonstrate how to recover domain-specific objects and relations, measure design properties, and use the resulting feedback to guide changes to the designs.

## 6 Conclusion and Future Work

We make spatial organization explicit across heterogeneous formats through a shared IR for space-syntax analysis. Across five diffusion backbones, empirically grounded post-training turns architectural knowledge into stronger design capabilities while preserving native mechanisms. Directly understanding designers' and users' intentions and needs, and iterating through simulated user feedback, remain beyond this theory-grounded scope. Future work should embed design-native representations and objectives in pretraining and investigate how design capability scales with data, model capacity, and compute. It should also validate the three-element workflow and harness through generator training and independent task evaluation across design domains.

## REPRODUCIBILITY STATEMENT

Code and curated experiment records are available at `https://anonymous.4open.science/r/GenAI4Design/README.md`.

The main paper specifies the frozen measurement and evaluation protocol, terminal feedback, and shared native-policy PPO objective. The appendix documents SSIO construction, data and metric support, five backbone policy bindings, independent checkpoint selection and final evaluation, and GSDiff-SSIO access completion. Five-block, condition-matched, connectivity, centrality, functional-profile, and reward-development analyses support the reported results.

## AI USE STATEMENT

The authors established the narrative, planned the exposition, and prepared an initial draft. AI tools then assisted with iterative checks of logical coherence and language polishing. Within the established code framework, they assisted with implementation details following the authors' algorithm and experiment design. They helped review the correctness and completeness of mathematical reasoning and notation and suggested additions and refinements. Building on an author-curated literature collection, they checked related-work coverage and identified recent relevant studies to supplement it.

## REFERENCES

Mohammadreza Aalaei, Melika Saadi, Morteza Rahbar, and Ahmad Ekhlassi. Architectural layout generation using a graph-constrained conditional generative adversarial network (GAN). *Automation in Construction*, 155:105053, 2023. doi: 10.1016/j.autcon.2023.105053.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=ShjMHfmPs0`.

Autodesk. Set up a static stress analysis, n.d. URL `https://help.autodesk.com/view/fusion360/ENU/?guid=SIM-SSA`. Official Autodesk Fusion documentation. Accessed 2026-09-15.

Vahid Azizi, Muhammad Usman, Samuel S. Sohn, Mathew Schwartz, Seonghyeon Moon, Petros Faloutsos, and Mubbasir Kapadia. The role of latent representations for design space exploration of floorplans. *Simulation*, 99(11):1167–1179, 2023. doi: 10.1177/00375497221115734.

In-Chang Baek, Sung-Hyun Kim, Seo-Young Lee, Dong-Hyeon Kim, and Kyung-Joong Kim. IPC-GRL: Language-Instructed Reinforcement Learning for Procedural Level Generation. In *2025 IEEE Conference on Games*, pp. 1–8, 2025. doi: 10.1109/COG64752.2025.11114105. URL `https://doi.org/10.1109/COG64752.2025.11114105`.

Devin J. Balkcom, Erik D. Demaine, Martin L. Demaine, John A. Ochsendorf, and Zhong You. Folding paper shopping bags. In Robert J. Lang (ed.), *Origami 4: Proceedings of the Fourth International Meeting of Origami Science, Math, and Education*, pp. 315–334. A K Peters, 2009. doi: 10.1201/b10653-33. URL `https://doi.org/10.1201/b10653-33`.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. AutomaTikZ: Text-guided synthesis of scientific vector graphics with TikZ. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=v3K5TVP8kZ`.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=YCWjhGrJFD`.

Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *Internet Mathematics*, 10(3–4):222–262, 2014. doi: 10.1080/15427951.2013.865686.

Martí Bosch, Maxence Locatelli, Perrine Hamel, Roy P. Remme, Jérôme Chenal, and Stéphane Joost. A spatially explicit approach to simulate urban heat mitigation with InVEST (v3.8.0). *Geoscientific Model Development*, 14(6):3521–3537, 2021. doi: 10.5194/gmd-14-3521-2021.

Louise P. Brown. TexGen. In Yordan Kyosev and Francois Boussu (eds.), *Advanced Weaving Technology*, pp. 253–291. Springer, Cham, 2022. doi: 10.1007/978-3-030-91515-5_6.

Nahyang Byun and Jaepil Choi. A typology of korean housing units: In search of spatial configuration. *Journal of Asian Architecture and Building Engineering*, 15(1):41–48, 2016. doi: 10.3130/jaabe.15.41.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. ArtiMuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15322, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cao_ArtiMuse_Fine-Grained_Image_Aesthetics_Assessment_with_Joint_Scoring_and_Expert-Level_CVPR_2026_paper.html.

Stuart K. Card, Thomas P. Moran, and Allen Newell. The Keystroke-Level Model for User Performance Time with Interactive Systems. *Communications of the ACM*, 23(7):396–410, 1980. doi: 10.1145/358886.358895. URL https://doi.org/10.1145/358886.358895.

Di Chen and Chang Shik Jang. A Comparative Study on the Visual Expression of Packaging Design Based on Generative AI: Focused on New Craft Beer Brands. *Journal of Cultural Product & Design*, pp. 55–67, 2026. URL https://kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART003358503. Issue 85.

Wei Chen and Faez Ahmed. MO-PaDGAN: Reparameterizing Engineering Designs for augmented multi-objective optimization. *Applied Soft Computing*, 113:107909, 2021. doi: 10.1016/j.asoc.2021.107909. URL https://doi.org/10.1016/j.asoc.2021.107909.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vmSEVL19f.

Shiqi Dai, Xuanyu Zhu, Naiqi Li, Tao Dai, and Zhi Wang. Procedural Level Generation with Diffusion Models from a Single Example. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10021–10029, 2024. doi: 10.1609/aaai.v38i9.28865. URL https://doi.org/10.1609/aaai.v38i9.28865.

Peter C. Dawson. Examining the impact of euro-canadian architecture on inuit families living in arctic canada. In *Proceedings of the Fourth International Space Syntax Symposium*, pp. 21.1–21.16, London, UK, 2003. URL https://www.spacesyntax.net/symposia-archive/SSS4/fullpapers/21Dawsonpaper1.pdf.

Adrienne Deganutti, Purvanshi Mehta, Simon Hadfield, and Andrew Gilbert. GDB-Reward: From evaluation metrics to training rewards for graphic design. *arXiv preprint arXiv:2609.02813*, 2026. URL https://arxiv.org/abs/2609.02813.

Victor Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 113–126. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-demo.11. URL https://aclanthology.org/2023.acl-demo.11/.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=m7p5O7zblY.

Joris Dormans. Adventures in level design: Generating missions and spaces for action adventure games. In *Proceedings of the 2010 Workshop on Procedural Content Generation in Games*. ACM, 2010. URL https://www.pcgworkshop.com/archive/dormans2010adventures.pdf.

Sam Earle, Maria Edwards, Ahmed Khalifa, Philip Bontrager, and Julian Togelius. Learning Controllable Content Generators. In *2021 IEEE Conference on Games*, pp. 1–9, 2021. doi: 10.1109/CoG52621.2021.9619159. URL https://doi.org/10.1109/CoG52621.2021.9619159.

Sam Earle, Zehua Jiang, and Julian Togelius. Scaling, Control and Generalization in Reinforcement Learning Level Generators. In *2024 IEEE Conference on Games*, pp. 1–8, 2024. doi: 10.1109/CoG60054.2024.10645598. URL https://doi.org/10.1109/CoG60054.2024.10645598.

Erincik Edgü and Alper Ünlü. Relation of domestic space preferences with space syntax parameters. In *Proceedings of the Fourth International Space Syntax Symposium*, pp. 82.1–82.16, London, UK, 2003. URL https://www.spacesyntax.net/symposia-archive/SSS4/fullpapers/82Edgu-Unlupaper.pdf.

Sara Eloy and Rosália Guerreiro. Transforming housing typologies. space syntax evaluation and shape grammar generation. *arq.urb*, (15):86–114, 2016. URL https://revistaarqurb.com.br/arqurb/article/view/233.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885, 2023.

Stanislava Fedorova. Generative Adversarial Networks for Urban Block Design. In *Proceedings of the Symposium on Simulation for Architecture and Urban Design*, 2021. URL https://simaud.org/proceedings/repo/2021-Papers/38-Fedorova.pdf.

FEFCO. FEFCO code: Design styles library for corrugated board products, 2022. URL https://www.fefco.org/download/file/fid/3022. 12th edition.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3256.

Jinding Gao, Chao Liang, Xiaocong Xu, Kang Chen, Xiaoping Liu, and Honghui Zhang. GenFusion: Crafting future urban building layouts via diffusion model and genetic algorithm. *Information Geography*, 1(1):100015, 2025. doi: 10.1016/j.infgeo.2025.100015. URL https://doi.org/10.1016/j.infgeo.2025.100015.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Interactive Story Visualization with Multiple Characters. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–10, 2023. doi: 10.1145/3610548.3618184. URL https://doi.org/10.1145/3610548.3618184.

Mertcan Güldilek, Mustafa Emre İlal, and Berk Ekici. Bridging GANs and reinforcement learning: Proximal policy optimization for enhanced design outputs. In *Transindividual Intelligence: Proceedings of the 7th International Conference on Computational Design and Robotic Fabrication (CDRF 2025)*, pp. 107–117, Singapore, 2026. Springer. doi: 10.1007/978-981-92-0615-5_10.

Su Guo, Guangce Liu, Haosen Yang, Jiepeng Wang, Cong Liu, Junqi Liu, Haibin Huang, Hongxun Yao, Chi Zhang, and Xuelong Li. ShotPlan: Cinematic Video Generation with Learnable Planning Token. *arXiv preprint arXiv:2607.17675*, 2026. URL https://arxiv.org/abs/2607.17675.

Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S. Davis, Vijay Mahadevan, and Abhinav Shrivastava. LayoutTransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1004–1014, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Gupta_LayoutTransformer_Layout_Generation_and_Completion_With_Self-Attention_ICCV_2021_paper.html.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

Gerald J. Hahn and William Q. Meeker. *Statistical Intervals: A Guide for Practitioners*. John Wiley & Sons, 1991. doi: 10.1002/9780470316771. URL https://doi.org/10.1002/9780470316771.

Julienne Hanson. *Decoding Homes and Houses*. Cambridge University Press, Cambridge, UK, 1998. doi: 10.1017/CBO9780511518294.

Feixiang He, Yanlong Huang, and He Wang. iPLAN: Interactive and procedural layout planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7793–7802, 2022.

Kai He, Kaixin Yao, Qixuan Zhang, Jingyi Yu, Lingjie Liu, and Lan Xu. DressCode: Autoregressively Sewing and Generating Garments from Text Guidance. *ACM Transactions on Graphics*, 43(4): 1–13, 2024. doi: 10.1145/3658147. URL https://doi.org/10.1145/3658147.

Lucy Helme, Christian Derix, and Åsmund Izaki. Spatial configuration: Semi-automatic methods for layout generation in practice. *The Journal of Space Syntax*, 5(1):35–49, 2014.

Shinjae Heo and Taigyoun Cho. Analysis of generative AI's floor plan images using visibility graphs. *Journal of the Korea Institute of Spatial Design*, 18(8):341–352, 2023. doi: 10.35216/kisd.2023.18.8.341.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

B. Hillier, A. Penn, J. Hanson, T. Grajewski, and J. Xu. Natural movement: Or, configuration and attraction in urban pedestrian movement. *Environment and Planning B: Planning and Design*, 20 (1):29–66, 1993. doi: 10.1068/b200029.

Bill Hillier. *Space Is the Machine: A Configurational Theory of Architecture*. Cambridge University Press, Cambridge, UK, 1996.

Bill Hillier and Julienne Hanson. *The Social Logic of Space*. Cambridge University Press, Cambridge, UK, 1984. doi: 10.1017/CBO9780511597237.

Shibo Hong, Xuhong Zhang, Tianyu Du, Sheng Cheng, Xun Wang, and Jianwei Yin. Cons2Plan: Vector floorplan generation from various conditions via a learning framework based on conditional diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3248–3256, 2024. doi: 10.1145/3664647.3680681.

Ruizhen Hu, Zeyu Huang, Yuhan Tang, Oliver Van Kaick, Hao Zhang, and Hui Huang. Graph2Plan: Learning floorplan generation from layout graphs. *ACM Transactions on Graphics*, 39(4):118:1–118:14, 2020. doi: 10.1145/3386569.3392391.

Sizhe Hu, Wenming Wu, Yuntao Wang, Benzhu Xu, and Liping Zheng. GSDiff: Synthesizing vector floorplans via geometry-enhanced structural graph generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17323–17332, 2025a. doi: 10.1609/aaai.v39i16.33904.

Zijing Hu, Fengda Zhang, Long Chen, Kun Kuang, Jiahui Li, Kaifeng Gao, Jun Xiao, Xin Wang, and Wenwu Zhu. Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23604–23614, 2025b.

Rong Huang, Haichuan Lin, Chuanzhang Chen, Kang Zhang, and Wei Zeng. PlantoGraphy: Incorporating Iterative Design Process into Generative Artificial Intelligence for Landscape Rendering. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2024. doi: 10.1145/3613904.3642824. URL https://doi.org/10.1145/3613904.3642824.

Robin Hunicke, Marc LeBlanc, and Robert Zubek. MDA: A formal approach to game design and game research, 2004. URL https://www.cs.northwestern.edu/~hunicke/MDA.pdf.

Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. LayoutDM: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10167–10176, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Inoue_LayoutDM_Discrete_Diffusion_Model_for_Controllable_Layout_Generation_CVPR_2023_paper.html.

International Safe Transit Association. Test procedures, n.d. URL https://www.ista.org/test_procedures.php. Official procedure-selection overview. Accessed 2026-09-15.

Bin Jiang, Christophe Claramunt, and Björn Klarqvist. Integration of space syntax into GIS for modelling urban spaces. *International Journal of Applied Earth Observation and Geoinformation*, 2(3–4):161–171, 2000. doi: 10.1016/S0303-2434(00)85010-2.

Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Huayu Deng, Zhongkai Wu, Vuksan Mijovic, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. LayoutFormer++: Conditional graphic layout generation via constraint serialization and decoding space restriction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18403–18412, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Jiang_LayoutFormer_Conditional_Graphic_Layout_Generation_via_Constraint_Serialization_and_Decoding_CVPR_2023_paper.html.

Halil Faruk Karagoz, Gulcin Baykal, Irem Arikan Eksi, and Gozde Unal. Textile Pattern Generation Using Diffusion Models. In *Proceedings of the International Textile and Fashion Congress*, 2023. URL https://ituyayinevi.itu.edu.tr/docs/librariesprovider59/bildiriler/%C4%B1tfc-2023-proceed%C4%B1ng-book_04-04.pdf.

Mohammad Sadil Khan, Sankalp Sinha, Talha Uddin Sheikh, Didier Stricker, Sk Aziz Ali, and Muhammad Zeshan Afzal. Text2CAD: Generating sequential CAD designs from beginner-to-expert level text prompts. In *Advances in Neural Information Processing Systems*, volume 37, pp. 7552–7579. Curran Associates, Inc., 2024. doi: 10.52202/079017-0242. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0e5b96f97c1813bb75f6c28532c2ecc7-Abstract-Conference.html.

Maria Korosteleva, Timur Levent Kesdogan, Fabian Kemper, Stephan Wenninger, Jasmin Koller, Yuhan Zhang, Mario Botsch, and Olga Sorkine-Hornung. GarmentCodeData: A dataset of 3D made-to-measure garments with sewing patterns, 2024. URL https://arxiv.org/abs/2405.17609.

Jianyu Lai, Sixiang Chen, Jialin Gao, Hengyu Shi, Zhongying Liu, Fuxiang Zhai, Junfeng Luo, Xiaoming Wei, Lujia Wang, and Lei Zhu. PosterReward: Unlocking accurate evaluation for high-quality graphic design generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7762–7772, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lai_PosterReward_Unlocking_Accurate_Evaluation_for_High-Quality_Graphic_Design_Generation_CVPR_2026_paper.html.

Jane M. Lamb and M. Jo Kallal. A conceptual framework for apparel design. *Clothing and Textiles Research Journal*, 10(2):42–47, 1992. doi: 10.1177/0887302X9201000207. URL https://journals.sagepub.com/doi/10.1177/0887302X9201000207.

Luis Lara, Aristides Milios, ZhiHao Luo, Aditya Sharma, Ge Ya Luo, Christopher Beckham, Florian Golemo, and Christopher Pal. Generative floor plan design with LLMs via reinforcement learning with verifiable rewards. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26612–26627, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1326. URL https://aclanthology.org/2026.findings-acl.1326/.

Chris Lattner and Vikram Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *Proceedings of the 2004 International Symposium on Code Generation and Optimization*. IEEE, 2004. doi: 10.1109/CGO.2004.1281665. URL https://llvm.org/pubs/2004-01-30-CGO-LLVM.html.

Chris Lattner, Mehdi Amini, Uday Bondhugula, Albert Cohen, Andy Davis, Jacques Pienaar, River Riddle, Tatiana Shpeisman, Nicolas Vasilache, and Oleksandr Zinenko. MLIR: Scaling compiler infrastructure for domain specific computation. In *2021 IEEE/ACM International Symposium on Code Generation and Optimization*, pp. 2–14. IEEE, 2021. doi: 10.1109/CGO51591.2021.9370308. URL https://doi.org/10.1109/CGO51591.2021.9370308.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, Ziyi He, Wei Wang, Dalin Li, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Yuxiang Chen, Yan Shu, Yanran Zhang, Yilei Chen, Yixian Xu, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Hongzhu Shi, Yi Wang, Bing Zhao, Hu Wei, Lin Qu, and Chenfei Wu. Qwen-Image-Bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026a. URL https://arxiv.org/abs/2605.28091.

Sha Li, Stefano Petrangeli, Yu Shen, and Xiang Chen. From pixels to policies: Reinforcing spatial reasoning in language models for content-aware layout design. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 1509–1518. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-industry.104. URL https://aclanthology.org/2026.acl-industry.104/.

Zhou Li, Ziyu Song, and Shaoming Lu. How to generate a museum floor plan: LoRA for diffusion models, or GANs? *Frontiers of Architectural Research*, 15(3):1040–1058, 2026c. doi: 10.1016/j.foar.2025.08.017.

Hongjin Lian, Jian Ma, Hongjie Chen, Jia Li, Ruizhen Hu, Yu-Kun Lai, and Kun Li. CG-Floor: Centroid-guided diffusion for large-scale floorplan generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18354–18363, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lian_CG-Floor_Centroid-Guided_Diffusion_for_Large-Scale_Floorplan_Generation_CVPR_2026_paper.html.

Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. IDEA-Bench: How far are generative models from professional designing? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18541–18551, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liang_IDEA-Bench_How_Far_are_Generative_Models_from_Professional_Designing_CVPR_2025_paper.html.

LIFULL Co., Ltd. LIFULL HOME'S Dataset. Informatics Research Data Repository, National Institute of Informatics, 2015. URL https://doi.org/10.32130/idr.6.0. Dataset.

Jiachen Liu, Yuan Xue, Jose Duarte, Krishnendra Shekhawat, Zihan Zhou, and Xiaolei Huang. End-to-end graph-constrained vectorized floorplan generation with panoptic refinement. In *European Conference on Computer Vision*, pp. 547–562. Springer, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems*, volume 38, pp. 40783–40818, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3a10c46572628d58cb44fb705f25cbbf-Abstract-Conference.html.

Shengqi Liu, Yuhao Cheng, Zhuo Chen, Xingyu Ren, Wenhan Zhu, Lincheng Li, Mengxiao Bi, Xiaokang Yang, and Yichao Yan. Multimodal Latent Diffusion Model for Complex Sewing Pattern Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17640–17650, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Multimodal_Latent_Diffusion_Model_for_Complex_Sewing_Pattern_Generation_ICCV_2025_paper.html.

Xiaochen Liu, Jiajia Peng, Zhiwen Lu, Yongxue Wang, and Feng Liu. Knit-FLUX: Simulation of Knitted Fabric Images Based on Low-Rank Adaptation of Diffusion Models. *Applied Sciences*, 15(16):8999, 2025c. doi: 10.3390/app15168999. URL https://www.mdpi.com/2076-3417/15/16/8999.

LLVM Project. LLVM language reference manual. Online technical documentation, 2026. URL https://llvm.org/docs/LangRef.html. Documentation snapshot accessed 2026-09-15.

Zhi Hao Luo, Luis Lara, Ge Ya Luo, Florian Golemo, Christopher Beckham, and Christopher Pal. DStruct2Design: Data and benchmarks for data structure driven generative floor plan design. *arXiv preprint arXiv:2407.15723*, 2024. URL https://arxiv.org/abs/2407.15723.

Ziniu Luo and Weixin Huang. FloorplanGAN: Vector residential floorplan adversarial generation. *Automation in Construction*, 142:104470, 2022. doi: 10.1016/j.autcon.2022.104470.

Jock Mackinlay. Automating the Design of Graphical Presentations of Relational Information. *ACM Transactions on Graphics*, 5(2):110–141, 1986. doi: 10.1145/22949.22950. URL https://doi.org/10.1145/22949.22950.

Yi Man, Changxiang Ding, Zihuan Zhang, and Ping Shu. Deep learning and computer vision for small residential space layout generation and quantitative evaluation. *Journal of Building Engineering*, 126:116144, 2026. doi: 10.1016/j.jobe.2026.116144.

William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190, Information Sciences Institute, University of Southern California, Marina del Rey, CA, 1987. URL https://www.sfu.ca/rst/pdfs/Mann_Thompson_1987.pdf.

George Mavrotas. Effective implementation of the $\varepsilon$-constraint method in Multi-Objective Mathematical Programming problems. *Applied Mathematics and Computation*, 213(2):455–465, 2009. doi: 10.1016/j.amc.2009.03.037.

Richard E. Mayer. Research-based principles for the design of instructional messages: The case of multimedia explanations. *Document Design*, 1(1):7–19, 1999. doi: 10.1075/dd.1.1.02may. URL https://www.benjamins.com/catalog/dd.1.1.02may.

François Mazé and Faez Ahmed. Diffusion Models Beat GANs on Topology Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9108–9116, 2023. doi: 10.1609/aaai.v37i8.26093. URL https://doi.org/10.1609/aaai.v37i8.26093.

W. M. Meier. *Space Group Patterns*. Number 14 in IUCr Teaching Pamphlets. University College Cardiff Press, 1984. URL https://www.iucr.org/__data/assets/pdf_file/0016/14371/14_annotated.pdf. Electronic edition, 2001.

Kiyohiro Nakayama, Jan Ackermann, Timur Levent Kesdogan, Yang Zheng, Maria Korosteleva, Olga Sorkine-Hornung, Leonidas J. Guibas, Guandao Yang, and Gordon Wetzstein. AIpparel: A Multimodal Foundation Model for Digital Garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8138–8149, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Nakayama_AIpparel_A_Multimodal_Foundation_Model_for_Digital_Garments_CVPR_2025_paper.html.

National Park Service. Wayside Exhibit Design. Harpers Ferry Center, 2019. URL https://www.nps.gov/subjects/hfc/wayside-exhibit-design.htm. Last updated December 18, 2019. Accessed September 15, 2026.

Nelson Nauata, Kai-Hung Chang, Chin-Yi Cheng, Greg Mori, and Yasutaka Furukawa. House-GAN: Relational generative adversarial networks for graph-constrained house layout generation. In *European Conference on Computer Vision*, pp. 162–177. Springer, 2020.

Nelson Nauata, Sepidehsadat Hosseini, Kai-Hung Chang, Hang Chu, Chin-Yi Cheng, and Yasutaka Furukawa. House-GAN++: Generative adversarial layout refinement network towards intelligent computational agent for professional architects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13632–13641, 2021.

Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. Learning Layouts for Single-Page Graphic Designs. *IEEE Transactions on Visualization and Computer Graphics*, 20(8):1200–1213, 2014. doi: 10.1109/TVCG.2014.48. URL https://doi.org/10.1109/TVCG.2014.48.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Stephen E. Palmer. Common region: A new principle of perceptual grouping. *Cognitive Psychology*, 24(3):436–447, 1992. doi: 10.1016/0010-0285(92)90014-S. URL https://www.sciencedirect.com/science/article/pii/001002859290014S.

Wamiq Para, Paul Guerrero, Tom Kelly, Leonidas J. Guibas, and Peter Wonka. Generative layout modeling using constraint graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6690–6700, 2021.

Keundeok Park, Semiha Ergan, and Chen Feng. Quality assessment of residential layout designs generated by relational generative adversarial networks (GANs). *Automation in Construction*, 158: 105243, 2024. doi: 10.1016/j.autcon.2023.105243.

John Peponis, Ruth Conroy Dalton, Jean Wineman, and Nick Dalton. Measuring the effects of layout upon visitors' spatial behaviors in open plan exhibition settings. *Environment and Planning B: Planning and Design*, 31(3):453–473, 2004. doi: 10.1068/b3041. URL https://doi.org/10.1068/b3041.

Quynh Phung, Long Mai, Fabian David Caba Heilbron, Feng Liu, Jia-Bin Huang, and Cusuh Ham. CineVerse: Consistent Keyframe Synthesis for Cinematic Scene Composition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2626–2636, 2026. doi: 10.1109/WACV61042.2026.00258. URL https://openaccess.thecvf.com/content/WACV2026/html/Phung_CineVerse_Consistent_Keyframe_Synthesis_for_Cinematic_Scene_Composition_WACV_2026_paper.html.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. URL https://arxiv.org/abs/2310.03739. Subsumed by arXiv:2407.08737.

Sizhong Qin, Chengyu He, Qiaoyun Chen, Sen Yang, Wenjie Liao, Yi Gu, and Xinzheng Lu. ChatHouseDiffusion: Prompt-guided generation and editing of floor plans. *arXiv preprint arXiv:2410.11908*, 2024. URL https://arxiv.org/abs/2410.11908.

Sizhong Qin, Ramon Elias Weber, and Xinzheng Lu. Tokenization allows multimodal large language models to understand, generate and edit architectural floor plans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10430–10440, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Qin_Tokenization_Allows_Multimodal_Large_Language_Models_to_Understand_Generate_and_CVPR_2026_paper.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Talya B. Rechavi. A room for living: Private and public aspects in the experience of the living room. *Journal of Environmental Psychology*, 29(1):133–143, 2009. doi: 10.1016/j.jenvp.2008.05.001.

Lyle Regenwetter, Akash Srivastava, Dan Gutfreund, and Faez Ahmed. Beyond Statistical Similarity: Rethinking Metrics for Deep Generative Models in Engineering Design. *Computer-Aided Design*, 165:103609, 2023. doi: 10.1016/j.cad.2023.103609. URL https://doi.org/10.1016/j.cad.2023.103609.

Allen Z. Ren, Justin Lidard, Lars L. Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mEpqHvbD2h.

Adrien Rougny, Vasundra Touré, Stuart Moodie, Irina Balaur, Tobias Czauderna, Hanna Borlinghaus, Ugur Dogrusoz, Alexander Mazein, Andreas Dräger, Michael L. Blinov, Alice Villéger, Robin Haw, Emek Demir, Huaiyu Mi, Anatoly Sorokin, Falk Schreiber, and Augustin Luna. Systems Biology Graphical Notation: Process Description language Level 1 Version 2.0. *Journal of Integrative Bioinformatics*, 16(2):20190022, 2019. doi: 10.1515/jib-2019-0022. URL https://sbgn.github.io/downloads/specifications/pd_level1_version2.pdf.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Mohammad Amin Shabani, Sepidehsadat Hosseini, and Yasutaka Furukawa. HouseDiffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5466–5475, 2023.

Kaede Shiohara and Toshihiko Yamasaki. Unified vector floorplan generation via markup representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39262–39271, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shiohara_Unified_Vector_Floorplan_Generation_via_Markup_Representation_CVPR_2026_paper.html.

Gillian Smith and Jim Whitehead. Analyzing the expressive range of a level generator. In *Proceedings of the 2010 Workshop on Procedural Content Generation in Games*, pp. 1–7, 2010. doi: 10.1145/1814256.1814260. URL https://www.pcgworkshop.com/archive/smith2010analyzing.pdf.

Tim J. Smith. The Attentional Theory of Cinematic Continuity. *Projections*, 6(1):1–27, 2012. doi: 10.3167/proj.2012.060102. URL https://doi.org/10.3167/proj.2012.060102.

John P. Snyder. Map projections: A working manual. Professional Paper 1395, U.S. Geological Survey, 1987. URL https://pubs.usgs.gov/publication/pp1395.

Nam P. Suh. Introduction to axiomatic design. Author-hosted chapter, Massachusetts Institute of Technology, 1996. URL https://web.mit.edu/2.882/www/chapter1/chapter1.htm. Chapter 1, version dated December 3, 1996.

Jiahui Sun, Wenming Wu, Ligang Liu, Wenjie Min, Gaofeng Zhang, and Liping Zheng. WallPlan: Synthesizing floorplans by learning to generate wall graphs. *ACM Transactions on Graphics*, 41 (4):92:1–92:14, 2022. doi: 10.1145/3528223.3530135.

Hao Tang, Zhenyu Zhang, Humphrey Shi, Bo Li, Ling Shao, Nicu Sebe, Radu Timofte, and Luc Van Gool. Graph transformer GANs for graph-constrained house generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2173–2182, 2023.

Qiang Tang, Liang Zheng, Yile Chen, Lina Yan, and Junzhang Chen. Artificial intelligence empowering museum space layout design: Insights from China. *PLOS ONE*, 19(11):e0310594, 2024a. doi: 10.1371/journal.pone.0310594. URL https://doi.org/10.1371/journal.pone.0310594.

Yi Tang, Congyao Xu, Yifan Xue, and Zhongfei Han. Generative Artificial Intelligence-Based Design of Functional Textile Structures and Their Applications in Artistic Expression. *Textile & Leather Review*, 9:4653–4677, 2026. doi: 10.31881/TLR.2026.4653. URL https://doi.org/10.31881/TLR.2026.4653.

Zecheng Tang, Chenfei Wu, Juntao Li, and Nan Duan. LayoutNUWA: Revealing the hidden layout expertise of large language models. In *International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/a6a1e4c756d700d9aedcc1896a7e6fb0-Abstract-Conference.html.

Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. ChartGPT: Leveraging LLMs to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, 31(3):1731–1745, 2025. doi: 10.1109/TVCG.2024.3368621. URL https://doi.org/10.1109/TVCG.2024.3368621.

Barbara Tversky, Julie Bauer Morrison, and Mireille Bétrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247–262, 2002. doi: 10.1006/ijhc.2002.1017. URL https://www.sciencedirect.com/science/article/pii/S1071581902910177.

Karl Ulrich. The role of product architecture in the manufacturing firm. *Research Policy*, 24:419–440, 1995. doi: 10.1016/0048-7333(94)00775-3. URL https://doi.org/10.1016/0048-7333(94)00775-3.

Abhinav Upadhyay, Alpana Dubey, Veenu Arora, Suma Mani Kuriakose, and Shaurya Agarawal. FLNet: Graph constrained floor layout generation. In *IEEE International Conference on Multimedia and Expo Workshops*, pp. 1–6, 2022. doi: 10.1109/ICMEW56448.2022.9859350.

Casper van Engelenburg, Fatemeh Mostafavi, Emanuel Kuhn, Yuntae Jeon, Michael Franzen, Matthias Standfest, Jan van Gemert, and Seyran Khademi. MSD: A benchmark dataset for floor plan generation of building complexes. In *Computer Vision – ECCV 2024*, volume 15116 of *Lecture Notes in Computer Science*, pp. 60–75, Cham, 2025. Springer. doi: 10.1007/978-3-031-73636-0_4.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Chenglong Wang, Bongshin Lee, Steven M. Drucker, Dan Marshall, and Jianfeng Gao. Data Formulator 2: Iterative creation of data visualizations, with AI transforming data along the way. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2025. doi: 10.1145/3706598.3713296. URL https://doi.org/10.1145/3706598.3713296.

Shidong Wang and Renato Pajarola. Eliminating rasterization: Direct vector floor plan generation with DiffPlanner. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):7906–7922, 2025. doi: 10.1109/TVCG.2025.3559682.

Shidong Wang and Renato Pajarola. Directly from alpha to omega: Controllable end-to-end vector floor plan generation. *IEEE Transactions on Visualization and Computer Graphics*, 32(7):5248–5261, July 2026. doi: 10.1109/TVCG.2026.3665422.

Shidong Wang, Wei Zeng, Xi Chen, Yu Ye, Yu Qiao, and Chi-Wing Fu. ActFloor-GAN: Activity-guided adversarial networks for human-centric floorplan design. *IEEE Transactions on Visualization and Computer Graphics*, 29(3):1610–1624, 2023. doi: 10.1109/TVCG.2021.3126478.

Jingxuan Wei, Cheng Tan, Qi Chen, Gaowei Wu, Siyuan Li, Zhangyang Gao, Linzhuang Sun, Bihui Yu, and Ruifeng Guo. From words to structured visuals: A benchmark and framework for text-to-diagram generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13315–13325, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wei_From_Words_to_Structured_Visuals_A_Benchmark_and_Framework_for_CVPR_2025_paper.html.

Jason Wu, Yi-Hao Peng, Xin Yue Amanda Li, Amanda Swearngin, Jeffrey P. Bigham, and Jeffrey Nichols. UIClip: A data-driven model for assessing user interface design. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–16. Association for Computing Machinery, 2024a. doi: 10.1145/3654777.3676408. URL https://doi.org/10.1145/3654777.3676408.

Jason Wu, Eldon Schoop, Alan Leung, Titus Barik, Jeffrey Bigham, and Jeffrey Nichols. UICoder: Finetuning large language models to generate user interface code through automated feedback. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7511–7525. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.naacl-long.417. URL https://aclanthology.org/2024.naacl-long.417/.

Rundi Wu, Chang Xiao, and Changxi Zheng. DeepCAD: A Deep Generative Network for Computer-Aided Design Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6772–6782, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Wu_DeepCAD_A_Deep_Generative_Network_for_Computer-Aided_Design_Models_ICCV_2021_paper.html.

Wenming Wu, Xiao-Ming Fu, Rui Tang, Yuhan Wang, Yu-Hao Qi, and Ligang Liu. Data-driven interior plan generation for residential buildings. *ACM Transactions on Graphics*, 38(6):234:1–234:12, 2019. doi: 10.1145/3355089.3356556.

Qu Xu, Jiajia Li, and Junsong Zhang. Rule and Data-Driven Automatic Exhibition Hall Space Planning and Layout Generation. *Journal of Computer-Aided Design & Computer Graphics*, 37(5):780–789, 2025. doi: 10.3724/SP.J.1089.2023-00439. URL https://doi.org/10.3724/SP.J.1089.2023-00439.

Xiang Xu, Joseph G. Lambourne, Pradeep Kumar Jayaraman, Zhengqing Wang, Karl D. D. Willis, and Yasutaka Furukawa. BrepGen: A B-rep Generative Diffusion Model with Structured Latent Geometry. *ACM Transactions on Graphics*, 43(4):1–14, 2024. doi: 10.1145/3658129. URL https://doi.org/10.1145/3658129.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024a.

Tao Yang, Yingmin Luo, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. PosterLLaVa: Constructing a unified multi-modal layout generator with LLM. *arXiv preprint arXiv:2406.02884*, 2024b. URL https://arxiv.org/abs/2406.02884.

Jun Yin, Pengyu Zeng, Haoyuan Sun, Yuqin Dai, Han Zheng, Miao Zhang, Yachao Zhang, and Shuai Lu. FloorPlan-LLaMa: Aligning architects' feedback and domain knowledge in architectural floor plan generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6640–6662, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.331. URL https://aclanthology.org/2025.acl-long.331/.

Pengyu Zeng, Wen Gao, Jun Yin, Pengjian Xu, and Shuai Lu. Residential floor plans: Multi-conditional automatic generation using diffusion models. *Automation in Construction*, 162:105374, 2024. doi: 10.1016/j.autcon.2024.105374.

Pengyu Zeng, Jun Yin, Yan Gao, Jizhizi Li, Zhanxiang Jin, and Shuai Lu. Comprehensive and dedicated metrics for evaluating AI-generated residential floor plans. *Buildings*, 15(10):1674, 2025a. doi: 10.3390/buildings15101674.

Pengyu Zeng, Jun Yin, Miao Zhang, Yuqin Dai, Jizhizi Li, ZhanXiang Jin, and Shuai Lu. CARD: Cross-modal agent framework for generative and editable residential design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9304–9319, Suzhou, China, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.473. URL https://aclanthology.org/2025.emnlp-main.473/.

Pengyu Zeng, Yuqin Dai, Jun Yin, Jing Zhong, Ziyang Han, Chaoyang Shi, ZhanXiang Jin, Maowei Jiang, Yuxing Han, and Shuai Lu. GreenPlanner: Practical floorplan layout generation via an energy-aware and function-feasible generative framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 8596–8606, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Zeng_GreenPlanner_Practical_Floorplan_Layout_Generation_via_an_Energy-Aware_and_Function-Feasible_CVPRF_2026_paper.html.

Changzhi Zhang, Liyuan Li, and Yibei Wang. Achieving architectural logic in automated floor plan design: A multi-stage double DQN framework from topological graph synthesis to rectangular layout generation. *Frontiers of Architectural Research*, 2026a. doi: 10.1016/j.foar.2026.04.010. Advance online publication.

Hang Zhang, Anton Savov, and Benjamin Dillenburger. MaskPLAN: Masked generative layout planning from partial input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8964–8973, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_MaskPLAN_Masked_Generative_Layout_Planning_from_Partial_Input_CVPR_2024_paper.html.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. STAGE: Storyboard-Anchored Generation for Cinematic Multi-shot Narrative. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 659–669, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_STAGE_Storyboard-Anchored_Generation_for_Cinematic_Multi-shot_Narrative_CVPR_2026_paper.html.

Hanyang Zhao, Haoxian Chen, Ji Zhang, David D. Yao, and Wenpin Tang. Scores as actions: A framework of fine-tuning diffusion models by continuous-time reinforcement learning. *arXiv preprint arXiv:2409.08400*, 2024. URL https://arxiv.org/abs/2409.08400.

Yulai Zhao, Masatoshi Uehara, Gabriele Scalia, Sun-Yuan Kung, Tommaso Biancalani, Sergey Levine, and Ehsan Hajiramezanali. Adding conditional control to diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=svp1EBA6hA.

ZeFeng Zhao, XianZhi Wang, MoRan Wang, XinYou Zhang, and Wei Song. Research on the application and evaluation of Mongolian saddle culture in kumis packaging design based on AIGC. *Humanities and Social Sciences Communications*, 13(1):649, 2026. doi: 10.1057/s41599-026-06948-z. URL https://doi.org/10.1057/s41599-026-06948-z.

Yu Zheng, Yuming Lin, Liang Zhao, Tinghai Wu, Depeng Jin, and Yong Li. Spatial planning of urban communities via deep reinforcement learning. *Nature Computational Science*, 3(9): 748–762, 2023. doi: 10.1038/s43588-023-00503-5. URL https://doi.org/10.1038/s43588-023-00503-5.

Zhaohua Zheng and Frank Petzold. Neural-guided room layout generation with bubble diagram constraints. *Automation in Construction*, 154:104962, 2023. doi: 10.1016/j.autcon.2023.104962.

Minjun Zhu, Zhen Lin, Yixuan Weng, Panzhong Lu, Qiujie Xie, Yifan Wei, Sifan Liu, Qiyao Sun, and Yue Zhang. AutoFigure: Generating and refining publication-ready scientific illustrations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5N3z9JQJKq.

Amir Abbas Ziaee, Georg Suter, and Laura Keiblinger. SFC-A68: A dataset for benchmarking space function and space access element classification methods in entire floors of multi-unit apartment buildings, 2024. URL https://doi.org/10.5281/zenodo.14245850.

Ziyang Zong, Guanying Chen, Zhaohuan Zhan, Fengcheng Yu, and Guang Tan. HouseTune: Two-stage floorplan generation with LLM assistance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(16):14059–14067, 2026. doi: 10.1609/aaai.v40i16.38417.

# A MEASUREMENT DEFINITIONS

SSIO uses common recovery, graph-construction, and aggregation rules across native formats; Appendix E.1 tests their effect on the findings.

## A.1 CANONICAL SSIO MEASUREMENT

Table 6: Notation used in the paper.

| Symbol | Meaning |
|---|---|
| *Design recovery and measurement* | |
| $c = (B, \mathcal{M}_c, G_c)$ | boundary, requested room multiset, optional relation graph |
| $\mathcal{A}_m$ | deterministic native-output normalization internal to SSIO |
| $\mathcal{C}_\phi$ | frozen access-completion map when native access is absent |
| $z_{p,m}$ | canonical measurement record for plan $p$ and method $m$ |
| $G_{p,m} = (V_{p,m}, E_{p,m})$ | room-labeled rectangular convex-atom access graph |
| $\mathcal{R}^+_{p,m}; V_j$ | retained rooms; retained atoms of room $j$ |
| $S_v; I_v, H_v$ | node centrality; Legacy HH integration and global harmonic centrality |
| $\bar{S}_j, \bar{S}_{p,k}, S_{p,g}$ | room, category, and domain scores |
| $\eta^{\text{eval}}, \eta^{\text{train}}_m$ | common evaluation settings and backbone-specific training-time measurement settings |
| *Functional organization and benchmark evaluation* | |
| $\mathcal{G}, \mathcal{H}_p$ | four functional domains and those with score support in plan $p$ |
| $\mathcal{K}_p, \mathcal{K}_{p,g}$ | categories with retained score support in plan $p$, and those within domain $g$ |
| $r^{\text{dom}}_{p,g}(S)$ | relative domain score within plan $p$ |
| $\widetilde{\mathbf{r}}(\mathcal{C}; S)$ | functional profile: four-domain median vector of cohort $\mathcal{C}$ |
| $\Delta_{\text{pub}}(p)$ | Public-domain margin within plan $p$ |
| IPD, PIA | population-profile and plan-anchor benchmark metrics |
| *Feedback and native-policy learning* | |
| $R_{\text{anchor}}, R_{\text{AreaGuard}}$ | terminal-feedback components |
| $\widehat{R}_i, \widehat{A}_i$ | quantile-clipped terminal reward and detached standardized trajectory advantage |
| $(m, r, i, d)$ | backbone, outer iteration, rollout, and execution-order decision indices |
| $(t^{(m)}_d, t^{(m)}_{d+1})$ | native reverse-time indices before and after stochastic decision $d$ |
| $s^{(m)}_{i,d}, a^{(m)}_{i,d}$ | native RL state and joint reverse action at decision $d$ |
| $\Omega^{(m)}_{i,d}$ | active stochastic coordinates in the joint policy density |
| $D_m, \mathcal{D}^{(m)}_i$ | number and ordered set $\{1, \dots, D_m\}$ of retained stochastic decisions |
| $\tau^{(m)}_i, \mathcal{P}_m$ | replayable trajectory and its outer-iteration rollout pool |
| $p_{\text{old},m}$ | iteration-start policy snapshot used to collect $\mathcal{P}_m$ |
| $\mathcal{B}_m$ | rollout–decision pairs included in PPO averaging |
| $N_m, M_m$ | rollouts per outer iteration and backbone-specific log-ratio bound |
| $L_m$ | padded room-slot capacity for room-tensor backbones |
| $\Theta_E$ | eligible post-training checkpoints |

For plan $p$ with condition $c_p$, source-specific SSIO normalization $\mathcal{A}_m$ exposes the ordered canonical record

$$z_{p,m} = \mathcal{A}_m(x^{\text{nat}}_{p,m}; c_p) = \big((Q_{p,m,j}, y_{p,m,j})_{j=1}^{R_{p,m}}, D^{\text{int}}_{p,m}\big), \tag{A.1}$$

where $Q_{p,m,j}$ is a walkable room region, $y_{p,m,j}$ its semantic label, and $D^{\text{int}}_{p,m}$ the interior access. For GSDiff-SSIO, the frozen completion map $\mathcal{C}_\phi$ supplies missing access relations before measurement.

Common evaluation settings $\eta^{\text{eval}}$ use a $256 \times 256$ lattice. Vector coordinates are mapped and clipped to its bounds before rasterization. The minimum retained-atom area is 50 pixels$^2$; the within-room continuity threshold is 1.5 pixels. Each generation attempt remains a separate observation, even with identical geometry.

For each room, SSIO repeatedly extracts the largest axis-aligned rectangle contained in its walkable mask. Rooms and candidate rectangles follow fixed source and raster traversal orders, and ties retain the first candidate in that order. Atoms from the same room are connected when they touch or fall within the continuity threshold; remaining within-room components are joined by their closest atom pair. An interior door adds an edge between the incident rooms' closest atoms. These rules produce a finite room-labeled rectangular convex-atom access graph and preserve room identity even when atom bounds coincide.

Table 7 gives the RPLAN semantic crosswalk. Bedroom subtypes merge only after room-level integration. Entrance, Storage, and unassigned structural regions may affect graph distances without becoming functional-profile coordinates; Section B.3 specifies their metric-support roles.

Table 7: Semantic crosswalk for the RPLAN benchmark.

| Canonical category | Source labels | Benchmark role |
|---|---|---|
| Living | Living | Public |
| Dining | Dining | Public |
| Bedroom | Master, Child, Second, Guest | Private |
| Study | Study | Private |
| Kitchen | Kitchen | Service |
| Bathroom | Bathroom | Service |
| Balcony | Balcony | Outdoor |
| Entrance | Entrance | PIA normalization only |
| Storage | Storage | PIA competitor |
| Wall-in/Unknown | Wall-in | graph/count only |

Fixed traversal, tie-breaking, semantic crosswalk, and finite-score conventions make measurement single-valued. A graph isomorphism preserving room membership and semantic labels preserves shortest-path distances and every supported room, category, domain, IPD, and PIA value. Scores depend on recovered structure rather than native encoding; door evidence and access completion determine that structure.

### A.2 Node Scores and Aggregation

For fixed $p, m$, write $G = (V, E)$ and suppress these indices on node and room scores. All node scores used here are finite and nonnegative on their scoring support. Let $C(v)$ be node $v$'s connected component, $n = |C(v)|$, and $d(v, u)$ the unweighted shortest-path distance. For $n > 2$, mean depth and relative asymmetry define component-local Legacy HH, finitely extended to degenerate cases:

$$\begin{aligned} \mathrm{MD}_v &= \frac{1}{n-1} \sum_{u \in C(v)\setminus\{v\}} d(v,u), \qquad \mathrm{RA}_v = \frac{2(\mathrm{MD}_v - 1)}{n-2}, \\ D_n &= \frac{2\{n[\log_2((n+2)/3) - 1] + 1\}}{(n-1)(n-2)}, \\ I_v &= \begin{cases} D_n/\mathrm{RA}_v, & n > 2 \text{ and } \mathrm{RA}_v > 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{A.2}$$

The zero-valued branch covers one- and two-node components and the universal-node case. Means use only retained, scored atoms and rooms.

Cross-corpus grounding uses global harmonic centrality (Boldi & Vigna, 2014), measuring whole-graph accessibility with normalization by $|V| - 1$:

$$H_v = \frac{1}{|V|-1} \sum_{\substack{u \in V\setminus\{v\} \\ d(v,u) < \infty}} \frac{1}{d(v,u)} \quad \text{for } |V| > 1, \tag{A.3}$$

with $H_v = 0$ for $|V| \le 1$. Unreachable nodes thus contribute zero, whereas universal nodes retain maximal finite centrality.

Room score $\bar{S}_j$ averages node scores $S_v$ equally over room $j$'s retained atoms; category score $\bar{S}_{p,k}$ averages retained room scores equally within category $k$. Domain score $S_{p,g}$ averages supported category scores within domain $g$ with scored-room-count weights. A category or domain is present for scoring only if it contains a scored room. Equation (1) defines the relative domain score and functional profile (four-domain vector of cohort medians); Section 3.2 defines the Public-domain margin, IPD, and PIA.

Table 8: Distinct summaries and learning signals derived from SSIO scores.

| Quantity | Aggregation or contrast | Role |
|---|---|---|
| Functional profile $\tilde{\mathbf{r}}$ | Within-plan relative domain scores, then cohort medians | Population organization; IPD |
| Public-domain margin $\Delta_{\mathrm{pub}}$ | Public minus strongest competing relative domain score | Individual domain hierarchy |
| PIA | Living category mean minus strongest eligible category mean; category-normalized | Individual RPLAN anchor |
| $R_{\mathrm{anchor}}$ | Largest Living room score minus largest non-Living room score; unnormalized | Post-training feedback |

Rescoring frozen graphs isolates sensitivity to the centrality definition. For $n > 2$, bounded HH uses $1/(1 + \mathrm{RA}_v/D_n)$, including one when $\mathrm{RA}_v = 0$; a connected pair scores one and an isolated node zero. An exclusion variant omits nodes in components of size at most two from aggregation. Room means, category weights, and domain support are recomputed from the remaining scored atoms and rooms, while the original Exact-8 acceptance counts remain fixed. Empty scoring support yields no statistic. Appendix E.1 reports sensitivity of the task-matched gap and post-training gains.

Records without retained atoms are not measurable. Statistics for requested categories, functional domains, or PIA competitors with empty support remain undefined, without imputation. Appendix B.3 specifies acceptance and support rules.

# B Data and Benchmark Protocol

## B.1 Data Roles

Table 9: Data roles in analysis and evaluation. Development, validation, calibration, and the endpoint pool below refer to the primary HD study; Table 11 gives the other backbones' training scales.

| Role | Size | Use |
|---|---|---|
| RPLAN real-corpus prior | 80,788 | cross-corpus functional regularity |
| LIFULL Living/Dining cohort | 19,677 / 19,541 | relative-domain-score / margin support |
| MSD Swiss corpus | 13,472 | cross-corpus functional regularity |
| SFC-A68 multi-country corpus | 270 | cross-corpus functional regularity |
| RPLAN exact-eight reference | 20,348 | frozen four-domain benchmark profile |
| SSPT-PPO rollout conditions | $\leq$ 7-room requests | post-training only |
| Functional validation | 2,000 | checkpoint eligibility and ordering |
| AreaGuard calibration | 1,000 | fixed area-scale statistics |
| Base quality calibration | $3 \times 5 \times 1,000$ | common FID/Compatibility ceilings; two protocols |
| Native validation | $1,000 \times 5$ | candidate FID and Compatibility |
| Native evaluation | $1,000 \times 5$ | post-selection FID and Compatibility |
| Shared evaluation suite | $5 \times 1,000$ | final benchmark and block stability |
| Matched endpoint pairs | 4,588 | condition-paired Base-to-SSPT-PPO analysis |

The five evaluation blocks have disjoint condition identifiers, absent from recorded rollout and checkpoint-selection pools. Each method makes one unreplaced attempt per condition; checkpoint selection precedes final-block scoring. The RPLAN reference is corpus-level rather than condition-paired; its exact-eight median PIA is 0.61718.

### B.2 Cross-Corpus Mapping

Across corpora, the crosswalk groups Living and Dining as Public, bedrooms and Study as Private, Kitchen and Bathroom as Service, and Balcony as Outdoor. Circulation and Storage enter access graphs but not domain aggregation. LIFULL closet/laundry maps to Storage and corridor to Entrance/circulation. Of 145,811 released records, 19,691 annotate Living or Dining; polygon reconstruction retains 19,677 for functional-profile estimation, including 19,541 supporting $\Delta_{\mathrm{pub}}$. MSD contributes 13,472 Swiss residential floorplans under the shared crosswalk. Apartment recovery from SFC-A68 access graphs retains 270 floorplans spanning 13 countries.

Appendix E.1 reports how centrality and connectivity choices affect the RPLAN and LIFULL Public-domain margins.

### B.3 Metric Support and Acceptance

Each functional-profile entry uses plans that contain its domain and have finite relative domain scores. For cohort $\mathcal{C}$, these plans form $\mathcal{C}_g = \{p \in \mathcal{C} : g \in \mathcal{H}_p\}$. All four domains must be represented to define the profile. Reporting IPD requires at least 100 such plans per domain overall and 20 per domain in each of five blocks; checkpoint validation requires only the overall minimum.

For PIA, normalization categories are Living, Bedroom, Kitchen, Bathroom, Dining, Study, Balcony, Entrance, and Storage. Eligible non-Living competitors are Bedroom, Kitchen, Bathroom, Dining, Study, Balcony, and Storage. Entrance normalizes scores but is not a competitor; Storage is. PIA requires Living and at least one eligible competitor.

All 20,348 reference plans contain Living; 429 (2.11%) also contain Dining. Living thus provides a consistently observed Public anchor. PIA retains Dining in both normalization and competition, whereas domain-level functional profiles and Public-domain margins aggregate Living and Dining as Public.

Let $e_i$ indicate whether attempt $i$ emits native output. After SSIO normalization and measurement, acceptance is

$$\begin{aligned} \mathbf{1}_{\mathrm{acc}}(i) &= \mathbf{1}\big[e_i = 1,\ z_i \text{ is measurable},\ N_i^{\mathrm{ret}} = 8,\ n_{i,\mathrm{Living}} > 0\big]\,, \\ \mathrm{Acc} &= \frac{\sum_i \mathbf{1}_{\mathrm{acc}}(i)}{N_{\mathrm{attempted}}}. \end{aligned} \tag{B.1}$$

Here $N_i^{\mathrm{ret}}$ counts retained spatial instances and $n_{i,\mathrm{Living}}$ retained Living rooms. Missing or malformed outputs remain in the attempted denominator. Acceptance requires exactly eight retained instances including Living, without requiring graph connectivity. IPD uses accepted outputs; PIA also requires an eligible competitor. Unsupported attempts receive no imputed functional score.

## C SSPT Realizations and Backbone Bindings

The selection-based realization motivates trajectory-space learning; the policy bindings below specify how shared feedback updates each backbone through its native decisions.

### C.1 Data-Space Selection

SSPT-Iter uses terminal reward to select training examples from a fixed real-data pool and $N$ generated candidates (Figure 8). The top $K$ supply ordinary diffusion updates where training targets are available; examples without targets are skipped without replacement. Ranking preserves source identities and resolves ties deterministically. Rejected candidates supply no negative feedback.

Appendix E.4.1 reports the controlled selection pilot and its contribution to the development of trajectory-space post-training.

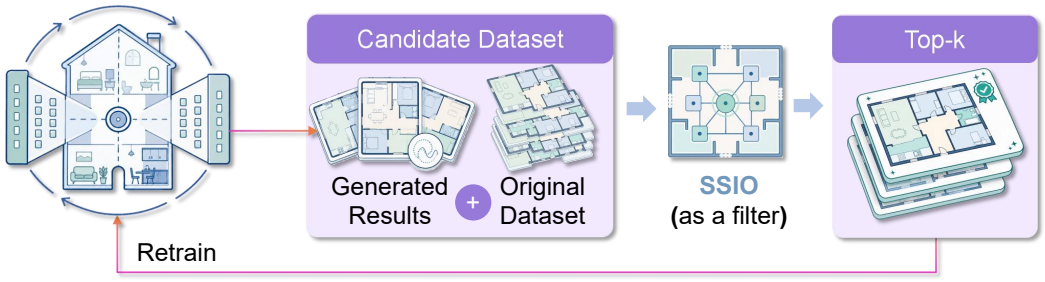



Figure 8: **SSPT-Iter: design feedback through data selection.** Generated candidates and a fixed real replay pool are ranked by SSIO scores. The top-$K$ selection supplies available targets for ordinary diffusion retraining before the next round.

## C.2 Backbone Policy Bindings

### C.2.1 Terminal Measurement and Reward Support

For fixed backbone $m$, terminal feedback uses $\eta_m^{\text{train}}$, with the retained-atom threshold in Table 11; final outputs are rescored under $\eta^{\text{eval}}$ at 50 pixels$^2$.

Training rewards and benchmark acceptance use different conditions. For a valid reward, $v_m(x) = 1$, every backbone requires successful native-output normalization. HD and ChatHD additionally require a scored Living space. CE2EPlan and DiffPlanner require finite scores for every native room and successful identification and realization of all official doors and their access openings, without a separate Living check. GSDiff requires finite measurements and consistent room counts. Empty maxima in $R_{\text{anchor}}$ are zero. Benchmark acceptance, by contrast, requires retained Living for every backbone.

For each backbone, a frozen set of Base outputs fixes $q_A$ at the 90th percentile of total retained-atom area and $s_A = \max\{\text{IQR}^{\text{base}}, 1 \text{ pixel}^2\}$, where IQR is the interquartile range of that area. For the primary HouseDiffusion run, $q_A = 22{,}134.4$ pixels$^2$ and $\text{IQR}^{\text{base}} = 3{,}621.5$ pixels$^2$; the selected anchor and AreaGuard weights are 8 and 4.

### C.2.2 Shared Trajectory Objective

At outer iteration $r$, the frozen snapshot $p_{\text{old},m} = p_{\theta_r,m}$ samples pool $\mathcal{P}_m = \{\tau_i^{(m)}\}_{i=1}^{N_m}$. Each trajectory records its condition, state–action–log-probability tuples, and terminal output:

$$\tau_i^{(m)} = \left( c_i, \left( s_{i,d}^{(m)}, a_{i,d}^{(m)}, \ell_{\text{old},i,d}^{(m)} \right)_{d \in \mathcal{D}_i^{(m)}}, x_{i,m}^{\text{nat}} \right). \tag{C.1}$$

The active-coordinate set $\Omega_{i,d}^{(m)}$ follows deterministically from the recorded state and fixed masking rule. The pool is replayed throughout the PPO epochs, then refreshed from the updated model.

For this fixed rollout pool, the finite empirical objective is

$$\widehat{J}_m(\theta) = \frac{1}{|\mathcal{B}_m|} \sum_{(i,d) \in \mathcal{B}_m} \min\left\{ \widetilde{\rho}_{i,d}^{(m)} \widehat{A}_i,\ \text{clip}\left( \widetilde{\rho}_{i,d}^{(m)}, 1-\epsilon, 1+\epsilon \right) \widehat{A}_i \right\}, \tag{C.2}$$

Training minimizes $-\widehat{J}_m$. Pool-level reward clipping, centering, and standardization are fixed before replay; the detached $\widehat{A}_i$ is shared by every recorded decision in trajectory $i$.

All bindings use a joint diagonal-Gaussian density over their active stochastic coordinates (Table 10):

$$\log p_{\theta,m}^{(d)}(a_{i,d}^{(m)} \mid s_{i,d}^{(m)}) = -\frac{1}{2} \sum_{j \in \Omega_{i,d}^{(m)}} \left[ \log(2\pi v_{m,d,j}) + \frac{(a_{i,d,j}^{(m)} - \mu_{\theta,m,d,j})^2}{v_{m,d,j}} \right], \tag{C.3}$$

where $\mu_{\theta,m,d}$ is produced by the native reverse kernel from $s_{i,d}^{(m)}$, and $v_{m,d}$ is its fixed reverse variance.

Table 10: Native policy bindings used by the five reported diffusion backbones.

| Backbone | Native variable | Active coordinates | Trainable component |
|---|---|---|---|
| HouseDiffusion | $2 \times 100$ coordinate tensor | all 200 coordinates | denoiser |
| CE2EPlan | $6 \times L_m$ room tensor $[\text{pres}, \text{type}, x, y, w, h]$ | unpadded $x, y, w, h$ coordinates | denoiser |
| DiffPlanner | $4 \times L_m$ room-box tensor $xyxy$ | unpadded room boxes | PartitioningDiff |
| ChatHD | $1 \times 64 \times 64$ semantic raster | interior raster pixels | suffix U-Net |
| GSDiff–SSIO | $53 \times 10$ corner/semantic tensor | joint reverse tensor | main diffusion model |

### C.2.3 Sampling Entry and Reverse-Time Support

All five backbones train on $D_m = 99$ stochastic decisions between 100 retained reverse-time points. The $(80, 20, 0, 0)$ schedule selects 80, 20, 0, and 0 points from four equal sections of the original 1,000-step axis; sampling traverses the retained points from $499 \to 0$.

HD, CE2EPlan, DiffPlanner, and GSDiff initialize this shortened chain with standard Gaussian noise, subject to native condition masks, without first running a high-noise prefix. ChatHD (ChatHouseDiffusion) instead enters through the frozen prefix described below. Each terminal output then undergoes SSIO normalization $\mathcal{A}_m$.

Final evaluation keeps the same respaced sampler for CE2EPlan, DiffPlanner, and GSDiff, and the same prefix–suffix sampler for ChatHD. HD is evaluated with its original full 1,000-step ancestral sampler, for both the benchmark and native-quality metrics; its lower-noise schedule determines post-training trajectories rather than replacing the final native sampler.

**Concrete reverse kernel.** Let $\kappa_0 < \cdots < \kappa_{99}$ be the original timesteps retained by $(80, 20, 0, 0)$. Specifically, $\kappa_e = \text{round}(249e/79)$ for $0 \le e < 80$, and $\kappa_{80+j} = 250 + \text{round}(249j/19)$ for $0 \le j < 20$. Sampling proceeds from effective position $e = 99$ to $e = 0$.

To relate these positions to the main-paper notation, suppress $i$ and $m$ and write $x_e := x^{(m)}_{i,\kappa_e}$. For every stochastic transition $e \to e - 1$, the execution index is $d = 100 - e$; hence $t^{(m)}_d = \kappa_e$ and $t^{(m)}_{d+1} = \kappa_{e-1}$. The transition $e = 1 \to 0$ is the last recorded action.

The deterministic $e = 0$ kernel uses rollout parameters $\theta_r$; its terminal state is deterministically emitted or decoded as $x^{\text{nat}}_{i,m}$. Outputs and rewards remain fixed during replay, and the PPO surrogate covers only recorded stochastic decisions.

For HouseDiffusion, the state contains $(x_e, \kappa_e, c)$. Write $a_e := a^{(m)}_{i,d} = x_{e-1}$ and $p^{(e)}_\theta := p^{(d)}_{\theta,m}$ for $d = 100 - e$. With fixed respaced variance $v_e > 0$, its policy is

$$
\begin{aligned}
p^{(e)}_\theta(a_e \mid x_e, c) &= \mathcal{N}\Big(\text{vec}(a_e); \text{vec}(\mu^{(e)}_\theta(x_e, c)), v_e I_{200}\Big), \\
\log p^{(e)}_\theta(a_e \mid x_e, c) &= -\frac{1}{2} \sum_{j=1}^{200} \left[ \log(2\pi v_e) + \frac{\big(a_{e,j} - [\mu^{(e)}_\theta(x_e, c)]_j\big)^2}{v_e} \right].
\end{aligned}
\tag{C.4}
$$

The primary HouseDiffusion experiment averages the surrogate over all 99 stochastic transitions. At $e < 32$, its binary branch replaces the continuous output and is hard-quantized before the policy mean is computed, giving zero parameter gradient. Nonzero gradients therefore arise from $e = 32, \ldots, 99$ through the shared Transformer and continuous-coordinate head; their retained original timesteps span 101 through 499.

### C.2.4 Backbone Instantiations

In CE2EPlan's type-conditioned mode, presence and category are restored from the condition after every transition. The likelihood covers only unpadded center/size coordinates $x, y, w, h$, updating the shared trunk, condition encoders, and $xywh$ output rows. Deterministic SSIO postprocessing derives adjacency from the generated boxes.

DiffPlanner fixes the upstream outputs of NodeDiff and AdjacencyDiff and trains PartitioningDiff. Deterministic alignment and decoration are applied during SSIO normalization.

ChatHD first executes 25 denoising diffusion implicit model (DDIM) transitions with $\eta = 0$, from $t = 999$ to $t = 499$, with a frozen Base U-Net. The ancestral suffix trains a separate U-Net copy through interior-pixel actions; terminal class quantization precedes SSIO normalization.

GSDiff trains its topology-conditioned main diffusion model through continuous corner coordinates. The topology encoder and terminal edge stage are frozen; hard semantic thresholds block direct gradients through the semantic channels, whose predicted values can still change as the diffusion model is updated. Its SSIO normalization applies $\mathcal{C}_\phi$ before convex-space graph construction.

Table 11 gives the training settings; all bindings use $\epsilon = 0.2$.

Table 11: Numerical settings of the backbone training bindings.

| Backbone | Iter.×rollouts/iter. | PPO epochs | LR | $M_m$ | Atom area (px$^2$) |
|---|---|---|---|---|---|
| HouseDiffusion | $20 \times 10{,}000$ | 2 | $5 \times 10^{-6}$ | 10 | 50 |
| CE2EPlan | $20 \times 128$ | 2 | $5 \times 10^{-6}$ | 10 | 16 |
| DiffPlanner | $60 \times 512$ | 4 | $10^{-6}$ | 10 | 16 |
| ChatHD | $5 \times 512$ | 2 | $5 \times 10^{-6}$ | 10 | 50 |
| GSDiff–SSIO | $10 \times 512$ | 2 | $5 \times 10^{-6}$ | 20 | 16 |

### C.3 Base-Referenced Quality Calibration and Model Selection

**Why use a Base reference?** SSPT seeks better functional organization while retaining the generation capabilities on which it builds. Frozen Base HD therefore supplies the native-quality reference. Because repeated evaluations of an unchanged generator vary, we use an upper prediction range rather than require every checkpoint to match a single Base score. FID measures discrepancy from real renderings; Compatibility measures disagreement with requested access conditions. We constrain their five-repeat means and optimize design within these limits, following the $\varepsilon$-constraint principle (Mavrotas, 2009). AreaGuard controls expansion during learning; the calibrated range determines which resulting checkpoints qualify.

**Calibration evidence.** We retrospectively calibrate the limits from three archived Base evaluation batches, without using post-trained design scores or final-test outcomes. Each batch contains five native evaluations of 1,000 exact-eight conditions. Batches A and B use the loader-default protocol and share the same saved real-image cohort; batch V uses the explicit native-validation condition list. We retain batch-specific means and variances, then construct a common ceiling covering these two development protocols rather than a separate requirement for each setting. Repeated seeded records of V are counted once; the code provides all source scores. The new limits replace the earlier empirical limits, and selection is reapplied to the archived validation candidates.

**What uncertainty should the limit cover?** Eligibility uses a mean of five evaluations, so calibration must target another five-repeat mean, not an individual run or the unknown population mean. Let $X_{b,j,r}$ denote repeat $r$ of metric $j \in \{\mathrm{FID}, \mathrm{Comp}\}$ in calibration batch $b \in \{A, B, V\}$. Write $\bar{X}_{b,j}$ for its mean and $s_{b,j}$ for its sample standard deviation, using denominator $n_b - 1$. Under an independent normal model for repeat scores within each batch, an independent new $m$-repeat Base mean $\bar{X}^{\mathrm{new}}_{b,j}$ satisfies the Student prediction relation (Hahn & Meeker, 1991)

$$\frac{\bar{X}^{\mathrm{new}}_{b,j} - \bar{X}_{b,j}}{s_{b,j}\sqrt{1/n_b + 1/m}} \sim t_{n_b - 1}, \qquad n_b = m = 5. \tag{C.5}$$

The two variance terms account for estimating Base performance and for sampling the new evaluation. A confidence bound containing only $1/n_b$ would omit the latter. Repeat-level normality and independence are modelling assumptions; five repeats do not establish their tail accuracy. The early batches do not record repeat seeds; V records seeds 0–4. The prediction target uses new independent sampling, whereas paired seeds remain useful when comparing candidate models.

**From uncertainty to numerical limits.** We choose a nominal 5% upper-tail error budget jointly over the three batches and two metrics. This controls false exclusion of unchanged Base at the calibration target, rather than defining an allowable percentage of model degradation. With $\alpha = 0.05$, $B = 3$, and $J = 2$, Bonferroni allocation gives

$$U_{b,j} = \bar{X}_{b,j} + t_{1-\alpha/(BJ),\, n_b-1}\, s_{b,j}\sqrt{1/n_b + 1/m},$$
$$\tau_j^{\mathrm{raw}} = \max_b U_{b,j}, \qquad \tau_j = \frac{\lceil 10\tau_j^{\mathrm{raw}} \rceil}{10}. \tag{C.6}$$

Taking the maximum makes one ceiling accommodate each observed development setting; it does not pool their distributions. The calibration batches need not be independent of one another: the union bound gives

$$\Pr\big(\exists\, b, j : \bar{X}_{b,j}^{\mathrm{new}} > \tau_j\big) \le \sum_{b,j} \Pr\big(\bar{X}_{b,j}^{\mathrm{new}} > U_{b,j}\big) \le \alpha \tag{C.7}$$

under the prediction model, over repeated calibration and new Base evaluations. The 5% budget is an explicit operating choice. Upward rounding to a one-decimal grid makes the limits easy to apply without narrowing their coverage; unrounded values are retained in Table 12. Here $t_{0.9916667,4} = 3.9607865$, yielding $\tau_{\mathrm{FID}}^{\mathrm{raw}} = 9.4243$ and $\tau_{\mathrm{Comp}}^{\mathrm{raw}} = 2.7822$, hence the adopted ceilings $\tau_{\mathrm{FID}} = 9.5$ and $\tau_{\mathrm{Comp}} = 2.8$.

Table 12: **Base-only calibration of the native-quality range.** Each entry uses five 1,000-condition repeats. Standard deviations $s$ use denominator four. The upper bounds jointly cover six future Base means under Eq. (C.6); bold entries determine the common ceilings.

| | FID | | | Compatibility | | |
|---|---|---|---|---|---|---|
| Base batch | Mean | $s$ | $U_{b,j}$ | Mean | $s$ | $U_{b,j}$ |
| A: loader-default | 8.5157 | 0.1670 | 8.9342 | 2.5240 | 0.0650 | 2.6869 |
| B: loader-default | 8.6177 | 0.3220 | **9.4243** | 2.5066 | 0.0207 | 2.5586 |
| V: native validation | 7.4739 | 0.0996 | 7.7234 | 2.6020 | 0.0719 | **2.7822** |
| Adopted upper limit | | 9.5 | | | 2.8 | |

**What does this requirement preserve?** The ceilings provide a shared Base-referenced operating range across the documented development settings. They do not require equality with the Base mean on every matched cohort. In particular, a checkpoint may have larger errors than its matched Base while remaining within this common range; final evaluation must report those differences as well as design gains. The statistical coverage concerns repeated unchanged-Base measurements, not a guarantee of perceptual equivalence for post-trained models. Different rendering, sample counts, or condition populations require recalibration. A smaller tail-error budget increases nominal simultaneous coverage by widening the range: $\alpha = 0.01, 0.05, 0.10$ gives rounded FID/Compatibility limits of 9.9/2.9, 9.5/2.8, and 9.3/2.8, respectively. Appendix E.4.6 examines the practical effect of these ceilings and alternative quality requirements.

**How is the model selected?** For the primary HouseDiffusion experiment, checkpoints must have at least 100 validation plans contributing to each functional-profile entry, at least one valid PIA score, and acceptance of at least 0.90. All five native-metric repeats must be finite, with mean FID (Heusel et al., 2017) at most $\tau_{\mathrm{FID}}$ and mean Compatibility at most $\tau_{\mathrm{Comp}}$. Among checkpoints meeting these requirements, lower IPD takes priority; the remaining quantities break ties. From this nonempty finite set $\Theta_E$, select

$$\theta^* = \operatorname*{arg\,min}^{\mathrm{lex}}_{\theta_j \in \Theta_E} \big(\mathrm{IPD}_{\mathrm{val}}(\theta_j), -\operatorname{med}\mathrm{PIA}_{\mathrm{val}}(\theta_j), \mathrm{FID}_{\mathrm{val}}(\theta_j), \mathrm{Comp}_{\mathrm{val}}(\theta_j), j\big), \tag{C.8}$$

where $\mathrm{FID}_{\mathrm{val}}$ and $\mathrm{Comp}_{\mathrm{val}}$ are the five-repeat means, and $j$ is a fixed candidate-order tie-break. Ranking uses unrounded values.

**What does final testing establish?** Reapplying selection with the calibrated ceilings recovers the originally tested checkpoint: AreaGuard-4 at iteration 20. Its independent final-test results are unchanged and report both design gains and native-quality changes relative to matched Base HD (Tables 2 and 21). Neither these outcomes nor candidate design rankings enter the Base-only calibration. Appendix E.4.6 separates this reanalysis from alternative validation operating points.

# D Comparison and Implementation Details

## D.1 Comparison Scope

The main benchmark table compares eight systems under the same one-attempt measurement protocol: CE2EPlan (Wang & Pajarola, 2026), DiffPlanner (Wang & Pajarola, 2025), House-GAN++ (Nauata et al., 2021), House-GAN (Nauata et al., 2020), Graph2Plan (Hu et al., 2020), FloorplanGAN (Luo & Huang, 2022), ChatHouseDiffusion (Qin et al., 2024), and HouseDiffusion. GSDiff (Hu et al., 2025a) is studied separately as GSDiff–SSIO, whose frozen access completion supplies relations required by the design measurement (Appendix E.6). Figure 6(a) retains all nine systems as a landscape of measured outputs; the completed system also remains in the five-backbone paired comparison.

CE2EPlan, DiffPlanner, House-GAN, and FloorplanGAN use checkpoints trained or retrained on the common fewer-than-eight-room split. HouseDiffusion, House-GAN++, ChatHouseDiffusion, and GSDiff use released checkpoints. Graph2Plan uses its released checkpoint with a reproduction of the released alignment procedure.

The released GSDiff checkpoint's pretraining split overlaps 4,241 of the 5,000 evaluation conditions. Its paired comparison therefore measures within-system post-training change under shared initialization and completion, rather than generalization to conditions unseen in pretraining.

Appendix E.5 reports the paired design results and accepted measurement support.

Figure 6(b) normalizes improvements from Base (B) to +SSPT-PPO (S) by each backbone's remaining reference gap. Here $\mathrm{PIA}_B$ and $\mathrm{PIA}_S$ are the corresponding cohort medians. With the exact-eight RPLAN median $\mathrm{PIA}_{\mathrm{ref}} = 0.61718$, it reports

$$g_{\mathrm{IPD}} = \frac{\mathrm{IPD}_B - \mathrm{IPD}_S}{\mathrm{IPD}_B}, \qquad g_{\mathrm{PIA}} = \frac{\mathrm{PIA}_S - \mathrm{PIA}_B}{\mathrm{PIA}_{\mathrm{ref}} - \mathrm{PIA}_B}. \tag{D.1}$$

## D.2 Native-Quality Evaluation

The HD quality evaluation follows its native automatic benchmark (Shabani et al., 2023). Generated and condition-matched real room and door polygons are rendered with the same semantic palette, black outlines, and white background at $256 \times 256$ pixels. FID compares their 2,048-dimensional Inception features using `pytorch-fid`. Its real reference is the native evaluation cohort, distinct from the 20,348-plan reference used for IPD.

Compatibility is the mean number of missing or extra edges between the requested access graph and the graph recovered from generated door–room overlaps, with an outside node for the exterior door. Room identities are fixed by the condition; the native modified graph-edit score counts edge disagreements rather than searching over node permutations. Smaller values indicate closer condition agreement, not stronger functional organization.

Checkpoint validation and final evaluation use separate fixed cohorts of 1,000 exact-eight conditions. Each checkpoint is sampled five times, with seeds 0–4 paired across checkpoints; means and population standard deviations summarize sampling variation. The final native-quality run uses a generation batch of 1,000 and an FID feature batch of 64. Native metrics evaluate the generated cohort directly, independently of RPLAN-SSI acceptance filtering. Both Base and the selected checkpoint use the original full 1,000-step HD ancestral sampler.

Figure 6(c) compares iteration-10 development configurations against the calibrated Base-reference ceilings; the Base marker supplies an independently evaluated reference, not a paired development comparison. Development uses the native loader's exact-eight cohort, five repeats, and generation batch 2,048, whereas final evaluation uses the explicit condition list and repeat seeds above. Com-

parisons among guard strengths use the development cohort; Table 2 provides the matched final Base–SSPT comparison.

# E Additional Experimental Results

The evidence follows the main argument: grounding and diagnostic robustness support Section 3.2; benchmark stability, matched comparisons, ablations, and portability support the corresponding experimental results; the GSDiff case separates access recovery, measurement attribution, and learning through a fixed interface.

## E.1 Grounding and Diagnostic Robustness

### E.1.1 Cross-Corpus Grounding

Do the functional regularities and corpus differences persist across centrality and connectivity choices? Table 13 compares the RPLAN and LIFULL Public-domain margins under the scoring definitions in Appendix A.

Table 13: Robustness of $\Delta_{\mathrm{pub}}$ in RPLAN and LIFULL across centrality and connectivity choices.

| Centrality | Cohort | RPLAN/LIFULL $N$ | RPLAN med. | LIFULL med. | Ratio |
|---|---|---|---|---|---|
| Legacy HH | all eligible | 80,788 / 19,541 | 0.7162 | 0.3052 | 42.61% |
| Legacy HH | connected only | 73,997 / 6,625 | 0.7062 | 0.3013 | 42.66% |
| Bounded HH, $n > 2$ | all eligible | 80,710 / 18,554 | 0.2602 | 0.1210 | 46.48% |
| Bounded HH, $n > 2$ | connected only | 73,997 / 6,590 | 0.2583 | 0.1268 | 49.08% |
| Global harmonic | all eligible | 80,788 / 19,541 | 0.2791 | 0.1086 | 38.90% |
| Global harmonic | connected only | 73,997 / 6,625 | 0.2759 | 0.1048 | 37.99% |

Every combination in Table 13 retains a positive median $\Delta_{\mathrm{pub}}$. Across these settings, the LIFULL median is 37.99–49.08% of the RPLAN median. The common direction supplies a design prior, while this persistent scale difference calls for task-matched numerical references.

### E.1.2 Task-Matched Gap and Centrality Sensitivity

To test whether the diagnosed gap and post-training response depend on one centrality definition, we rescore the frozen spatial graphs under four settings. Each row evaluates Eq. (2) with the declared $S_v$ and its aggregates in place of $I_v$ and its aggregates, including the rescored reference. The official benchmark uses Legacy HH.

Table 14: Centrality sensitivity of the task-matched functional gap and SSPT-PPO gains.

| Centrality | Public-margin deficit | IPD Base→+ | median PIA Base→+ |
|---|---|---|---|
| Legacy HH | 46.27% | 0.1111→0.0811 | 0.2800→0.3974 |
| Bounded HH | 38.68% | 0.0368→0.0230 | 0.1060→0.1573 |
| Bounded HH, $n > 2$ | 38.54% | 0.0376→0.0223 | 0.1060→0.1576 |
| Global harmonic | 42.36% | 0.0503→0.0178 | 0.1186→0.2022 |

Every rescore in Table 14 starts from the same exact-eight pools: 20,348 RPLAN plans, 4,723 HD Base outputs, and 4,825 +SSPT-PPO outputs. The Public-margin deficit is the difference between the RPLAN and Base medians of $\Delta_{\mathrm{pub}}$, divided by the RPLAN median. After small-component exclusion, 20,338 RPLAN plans retain the required measurement support. The diagnosed gap and both post-training gains persist across all four centrality definitions, extending beyond the Legacy HH score used for learning.

### E.2 Benchmark and Post-Training Stability

Table 15 tests whether the pooled findings persist across five disjoint evaluation blocks, covering the eight-system benchmark and primary HD post-training result. GSDiff–SSIO block statistics accompany its separate case in Appendix E.6.

Table 15: Mean and population standard deviation across five evaluation blocks for the benchmark and primary post-training result. The main table pools all five blocks.

| | Configurational metrics | | Protocol outcome |
|---|---|---|---|
| Method | Block IPD ↓ | Block median PIA ↑ | Accepted |
| CE2EPlan | 0.1916±0.0020 | 0.0745±0.0114 | 3,324/5,000 |
| DiffPlanner | 0.1806±0.0028 | 0.1119±0.0080 | 2,956/5,000 |
| House-GAN++ | 0.1611±0.0034 | 0.1599±0.0100 | 2,036/5,000 |
| House-GAN | 0.1371±0.0047 | 0.2616±0.0080 | 1,803/5,000 |
| Graph2Plan | 0.1074±0.0018 | 0.2899±0.0095 | 3,582/5,000 |
| FloorplanGAN | 0.0983±0.0084 | 0.2288±0.0079 | 4,574/5,000 |
| ChatHouseDiffusion | 0.0845±0.0046 | 0.3298±0.0171 | 1,561/5,000 |
| HouseDiffusion Base | 0.1112±0.0009 | 0.2804±0.0027 | 4,723/5,000 |
| HouseDiffusion+SSPT-PPO | **0.0813±0.0018** | **0.3991±0.0098** | **4,825/5,000** |

The primary Base-to-SSPT-PPO gains are stable across the five evaluation blocks (Table 15). For HD, the mean block IPD changes from 0.1112 to 0.0813 and the mean block median PIA from 0.2804 to 0.3991. The corresponding block standard deviations are substantially smaller than these changes, supporting improvement across disjoint condition blocks.

### E.3 Matched Comparisons and Design Interpretation

#### E.3.1 Condition-Matched Sensitivity

To separate design improvement from changes in measurement coverage, Table 16 recomputes population and plan-level statistics on the 4,588 of 5,000 conditions jointly accepted by the Base and +SSPT-PPO arms. Both metrics improve on this identical condition subset.

Table 16: Condition-matched sensitivity on 4,588 jointly accepted outputs.

| Statistic | Base | +SSPT-PPO |
|---|---|---|
| median PIA [25th, 75th percentiles] | 0.2809 [0.1306, 0.3937] | 0.3968 [0.2870, 0.4916] |
| $\Pr[\text{PIA} > 10^{-6}]$ | 87.90% | 98.10% |
| IPD on joint subset | 0.1112 | 0.0812 |

Design improvement extends to most matched outputs: the paired $\Delta\text{PIA}$ has median 0.1164 and is positive for 3,231 of 4,588 conditions (70.42%). With 10,000 within-block bootstrap replicates of these jointly accepted pairs, its median has 95% interval $[0.1082, 0.1257]$; the improvement fraction has Wilson interval $[69.09\%, 71.73\%]$. The IPD bootstrap instead resamples all attempted conditions within blocks, recomputing each arm's acceptance and four population medians; its 95% interval for the reduction is $[25.2\%, 28.6\%]$.

#### E.3.2 Connectivity Sensitivity

Table 17: Single-component sensitivity on the five-block benchmark.

| Method | Conn./accepted | Conn./attempted | IPD | median PIA |
|---|---|---|---|---|
| Base | 4,223/4,723 (89.41%) | 4,223/5,000 (84.46%) | 0.1124 | 0.2742 |
| +SSPT-PPO | 4,311/4,825 (89.35%) | 4,311/5,000 (86.22%) | 0.0833 | 0.3883 |

Among connected layouts in each arm, IPD falls by 25.96% and median PIA rises by 41.61% (Table 17). A stricter comparison retains 3,914 condition-matched pairs connected in both arms: 70.49% improve in PIA, with median paired change $+0.1119$. Gains within this identical, already-connected subset show that post-training improves functional organization beyond repairing disconnected graphs.

#### E.3.3 Functional Profiles

Table 18: Median relative-integration profile coordinates.

| Source | Public | Private | Service | Outdoor |
|---|---|---|---|---|
| RPLAN reference | 1.5548 | 0.8772 | 0.7860 | 0.7445 |
| HouseDiffusion Base | 1.3465 | 0.9565 | 0.9210 | 0.7664 |
| HouseDiffusion+SSPT-PPO | 1.4044 | 0.9068 | 0.8686 | 0.8064 |

Post-training moves Public, Private, and Service toward the RPLAN reference, extending the effects of anchor-directed feedback across several functional domains. Outdoor moves away while the overall IPD decreases. The full profile reveals how relative integration changes across functional domains, including differences that an aggregate score alone conceals.

### E.4 Ablations and Method Development

The selection pilot explains the move from data-space ranking to trajectory feedback. Three controlled HD ablations then test reverse-time support, reward identity, and area-guard strength; independent validation selects the final checkpoint.

#### E.4.1 Data-Space Selection Pilot

The controlled pilot uses three rounds, $N = 10{,}000$ generated candidates per round, $K = 4{,}000$, and 80,788 real replay examples. All 12,000 selections come from the real pool; 141 selected examples per round lack a training target, leaving 3,859 update targets per round. None of the 30,000 generated candidates therefore contributes to learning. Ranking against real examples can exclude the very outputs that need improvement; trajectory-space post-training instead uses their measured rewards to supply positive or negative advantages.

#### E.4.2 Reverse-Time Support Ablation

Where in denoising should design feedback act? Table 3 compares three HouseDiffusion reverse-time supports under the same profile-distance reward and a fixed budget of 100 timesteps from the original 1,000-step axis: $(80, 20, 0, 0)$ for lower noise, $(25, 25, 25, 25)$ for uniform full-horizon coverage, and $(0, 0, 20, 80)$ for higher noise. Entries specify counts in four equal sections ordered from low to high noise. Each arm uses 5,000 trajectories per iteration, seed 0, two PPO epochs, microbatch size 32, and gradient accumulation over four microbatches. The comparison evaluates iteration 3 on 1,000 exact-eight conditions per arm.

Lower-noise support gives lower IPD and higher median PIA than uniform full-horizon or higher-noise support (Table 3). Thus, the same design feedback yields different gains depending on where updates act in HD's reverse process.

#### E.4.3 Reward Identity

The reward comparison tests whether public anchoring, category-profile alignment, or their combination provides more effective design feedback. Candidates were evaluated at iteration 10 on a 5,000-condition development pool distinct from final evaluation, using 10,000 rollout trajectories per iteration and the same lower-noise support. Anchor is Eq. (3) with $\lambda = 0$. Category alignment uses the negative mean absolute distance between supported within-plan category coordinates and their RPLAN medians. Living/MAD standardizes Living against the median and robust scale of present non-Living categories. Anchor+category alignment sums the first two signals. Undefined measurement support receives the fixed failure value rather than entering a partial average.

Table 19: Reward identity on the development pool at iteration 10.

| | Configurational metrics | | Protocol outcome |
|---|---|---|---|
| Reward | IPD ↓ | median PIA ↑ | Accepted |
| Anchor | 0.0830 | 0.3912 | 4753/5000 |
| Category alignment | 0.0828 | 0.3903 | 4782/5000 |
| Living/MAD | 0.0927 | 0.3554 | 4747/5000 |
| Anchor+category alignment | **0.0807** | **0.4027** | 4764/5000 |

Anchor and category alignment perform similarly, both ahead of Living/MAD (Table 19). Combining them further improves both design scores but worsens FID from the matched Base value of $7.474 \pm 0.089$ to $21.470 \pm 0.244$, exceeding the limit of 9.5. Stronger design feedback alone therefore does not preserve visual quality. The next comparison holds Anchor fixed to test whether controlling area expansion improves this trade-off.

### E.4.4 Area-Guard Strength

**What does the area guard contribute?** Holding Anchor and iteration 10 fixed isolates AreaGuard (Table 4, Figure 9). Weight 2 reduces FID from 19.55 to 8.91 and Compatibility from 2.583 to 2.118 while slightly improving both design metrics. Weight 4 further reduces native-metric errors but gives weaker design performance than weight 2. The guard therefore changes the attainable quality–design trade-off, rather than merely moving a checkpoint across a gate. Independent validation then selects guard strength and duration jointly under common quality budgets.

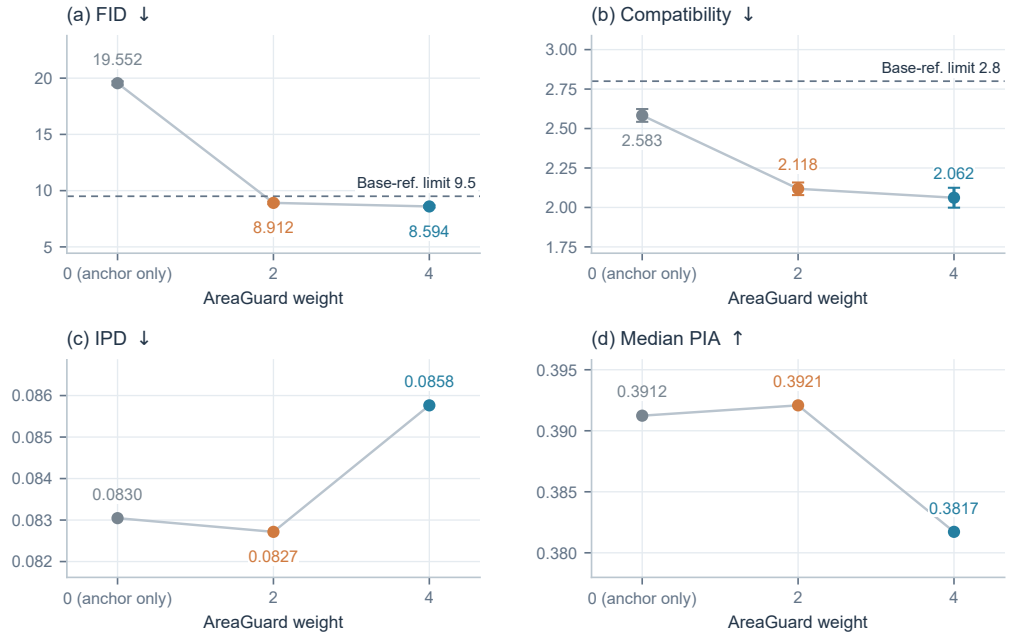



Figure 9: **AreaGuard strength at a fixed training duration.** Anchor and iteration 10 are held fixed. Weight 2 improves all four plotted metrics relative to no guard; weight 4 further favors native quality at some design-score cost. Design metrics use 5,000 development conditions; native metrics use five repeats of 1,000 conditions. Error bars show repeat standard deviations; dashed lines mark the calibrated Base-reference ceilings.

### E.4.5 Interpreting the Base-Reference Range

What does the calibrated FID ceiling allow in concrete outputs? We examine saved renderings independently of design-score ranking, comparing foreground and axis-aligned bounding-box areas with those of condition-matched real images. The two loader-default Base cohorts have mean per-image foreground-area ratios of 0.954 and 0.951; Anchor-only outputs reach 1.380. AreaGuard-2 and -4 reduce this ratio to 1.028 and 0.994. Foreground touches the outermost image border in 0.1% of Base renderings and 5.9% of Anchor renderings. These image-space measurements connect the FID cost to expansion; they describe rendered geometry, not physical floor area.

To examine intermediate changes, we replace selected Base images with the corresponding same-condition Anchor images while holding all 1,000 real reference images fixed. Two Base cohorts, three fixed nested substitution orders, and replacement fractions of 0, 1, 2, 5, 10, 15, 20, 25, 50, and 100% give 60 evaluated compositions. Every FID is recomputed from the actual 2,048-dimensional features, not interpolated between endpoint scores; recomputed saved endpoints differ from their archived FIDs by less than 0.01. Figure 10 focuses on the 0–25% range around the ceiling. Each composition supplies one FID; the substitution orders are not independent generation repeats and do not recalibrate the five-repeat limits.

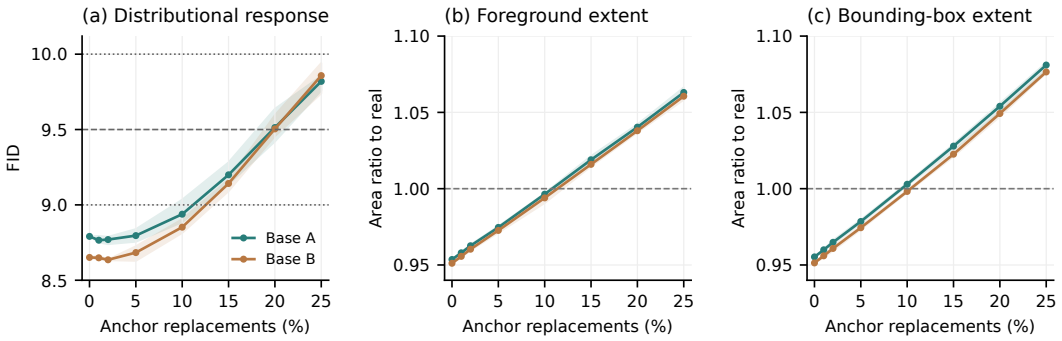



Figure 10: **What changes near the Base-referenced FID ceiling?** Replacing Base outputs with same-condition Anchor outputs increases geometric extent and eventually FID. Curves show means across three substitution orders; bands span their minima and maxima. Dashed lines mark FID 9.5 and unit geometric ratios; dotted FID lines show alternative ceilings. Ratios are averaged per-image ratios to the matched real rendering.

A limit of 8.5 excludes both early Base five-repeat FID means. In the controlled compositions, all six 15%-replacement cohorts have FID 9.091–9.281 while the mean foreground-area and bounding-box-area ratios are closer to one than Base: their single-cohort FIDs exceed 9.0 despite their closer geometric extent. Compositions below FID 9.5 include some overshoot, whereas those below 10.0 include larger departures. The largest mean bounding-box-area ratios among compositions below 9.0, 9.5, and 10.0 are 1.003, 1.056, and 1.083, respectively. These comparisons interpret the operating range's geometric consequences; they neither replace five-repeat eligibility checks nor fit a new threshold. Compatibility separately constrains requested access relations and cannot be inferred from rendered-area measurements.

### E.4.6 Guard Strength and Training Duration

We reapply Eq. (C.8) to nine archived checkpoints, each evaluated on 2,000 functional-validation conditions and five native-quality repeats of 1,000 conditions. Four satisfy the calibrated FID/Compatibility ceilings of 9.5/2.8 (Table 20). Among them, AreaGuard-4 at iteration 20 has the lowest IPD. The recalibration leaves both the eligible set and selected checkpoint unchanged: candidates newly below the Compatibility ceiling still exceed the FID ceiling. The independent final test therefore retains the same model and outputs.

Table 20: **Validation separates eligibility from design-based selection.** Native metrics are five-repeat means; acceptance is a fraction. All candidates satisfy support and acceptance requirements. Status applies the Base-calibrated FID/Compatibility ceilings of 9.5/2.8; the selected row is bold.

| Guard | Iter. | IPD ↓ | median PIA ↑ | Acc. | FID ↓ | Comp. ↓ | Status |
|---|---|---|---|---|---|---|---|
| 2 | 10 | 0.0842 | 0.3934 | 0.9665 | 8.594 | 2.210 | Eligible |
| 2 | 15 | 0.0812 | 0.4031 | 0.9610 | 10.061 | 2.430 | Excluded |
| 2 | 19 | 0.0778 | 0.4067 | 0.9560 | 12.330 | 2.640 | Excluded |
| 2 | 22 | 0.0811 | 0.4050 | 0.9510 | 13.557 | 2.787 | Excluded |
| 2 | 25 | 0.0822 | 0.4040 | 0.9445 | 15.084 | 2.967 | Excluded |
| 2 | 30 | 0.0835 | 0.3913 | 0.9210 | 19.145 | 3.343 | Excluded |
| 4 | 10 | 0.0843 | 0.3871 | 0.9705 | 8.176 | 2.184 | Eligible |
| 4 | 15 | 0.0839 | 0.3926 | 0.9690 | 8.462 | 2.331 | Eligible |
| **4** | **20** | **0.0835** | **0.3948** | **0.9630** | **9.359** | **2.581** | **Selected** |

**How does training duration change the trade-off?** With AreaGuard-2, validation IPD improves through iteration 19, while FID exceeds its limit at iteration 15; later checkpoints worsen on both design and native metrics. With weight 4, all three checkpoints satisfy the quality limits and both design metrics improve through iteration 20 (Figure 11). From iterations 15 to 20, IPD falls from 0.08390 to 0.08352 and median PIA rises from 0.39255 to 0.39484, while FID increases from 8.462 to 9.359 and Compatibility from 2.331 to 2.581. Iteration 15 thus retains better native quality with slightly weaker design scores. Iteration 20 is selected because the rule prioritizes design performance among models satisfying both quality limits.

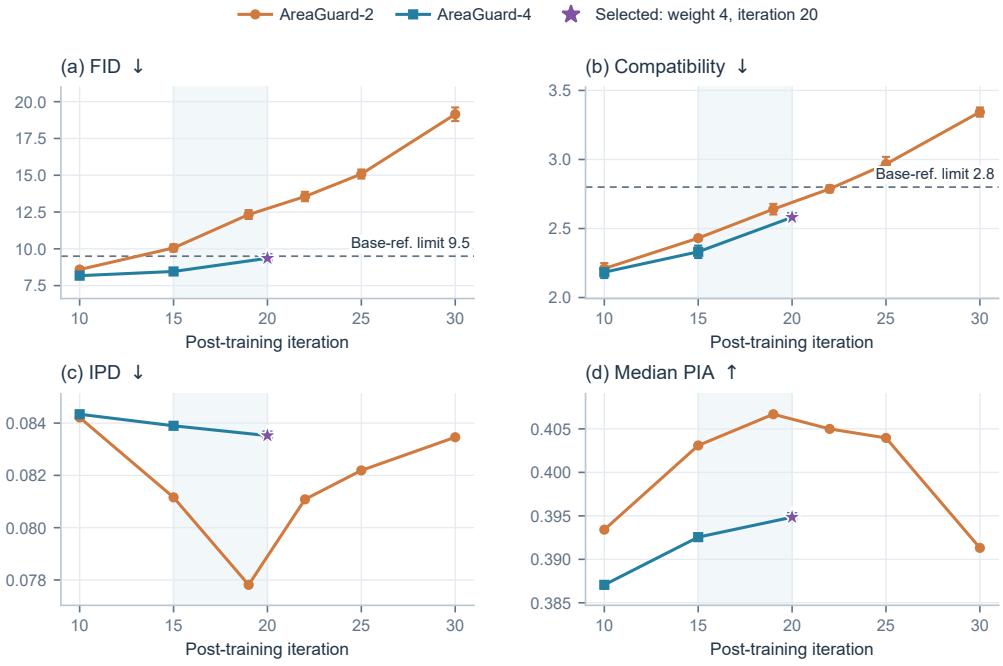



Figure 11: **Training duration and the quality–design trade-off.** Lines connect the nine measured validation checkpoints in Table 20; native-metric error bars show repeat standard deviations. Shading marks iterations 15–20, where weight 4 yields smaller additional design gains at increasing native cost. The star identifies the selected checkpoint. Dashed lines mark the Base-calibrated ceilings.

**How sensitive is selection to the quality requirement?** Figure 12 varies the ceilings while holding the measured candidates and ranking fixed. Tightening FID to 9.0, or Compatibility to 2.5, selects weight 4 at iteration 15. Relaxing FID to 10.1 with Compatibility at 2.8 selects weight 2 at iteration 15, offering better design scores at higher FID. The adopted 9.5/2.8 limits instead select weight

4 at iteration 20. Applying the unrounded Base-derived ceilings, 9.4243/2.7822, gives the same selection, so decimal rounding does not determine the result. These validation comparisons show the consequence of changing the quality requirement; they do not supply its calibration evidence. Table 2 reports independent final testing of the selected checkpoint.

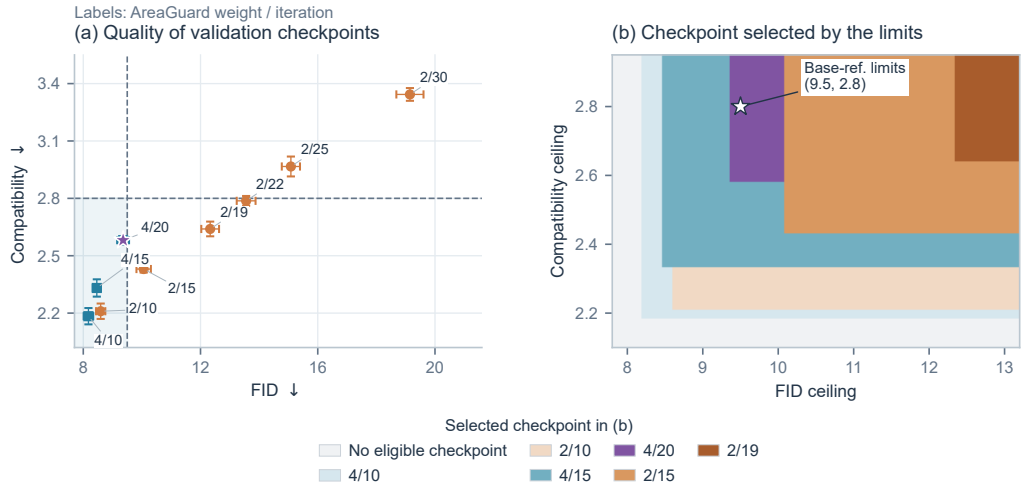



Figure 12: **Checkpoint selection under different quality requirements.** (a) Validation FID and Compatibility; shading marks the admissible region and error bars show repeat standard deviations. (b) Selection with the candidate set and Eq. (C.8) fixed; labels denote guard weight/iteration. The star marks the Base-calibrated ceilings, 9.5/2.8, selecting weight 4 at iteration 20. Stricter ceilings can select iteration 15.

### E.5 Cross-Backbone Results and Measurement Coverage

Does the shared design objective improve generators with different native mechanisms? Every portability pair uses the same 5,000 conditions for its Base and +SSPT-PPO arms. Table 21 reports both design metrics and accepted support under each backbone's fixed measurement binding. HD supplies the detailed native-quality, ablation, and case-study evidence; the other four pairs test portability.

Table 21: Full paired portability results, including Exact-8 acceptance.

| Backbone | Variant | IPD ↓ | median PIA ↑ | Accepted |
|---|---|---|---|---|
| HouseDiffusion | Base | 0.1111 | 0.2800 | 4,723/5,000 |
| | +SSPT-PPO | **0.0811** | **0.3974** | 4,825/5,000 |
| CE2EPlan | Base | 0.1914 | 0.0726 | 3,324/5,000 |
| | +SSPT-PPO | **0.1352** | **0.2441** | 2,956/5,000 |
| DiffPlanner | Base | 0.1803 | 0.1113 | 2,956/5,000 |
| | +SSPT-PPO | **0.1555** | **0.1721** | 3,227/5,000 |
| ChatHD | Base | 0.0866 | 0.3295 | 1,561/5,000 |
| | +SSPT-PPO | **0.0838** | **0.3336** | 1,556/5,000 |
| GSDiff–SSIO | Base | 0.0386 | 0.5766 | 3,186/5,000 |
| | +SSPT-PPO | **0.0360** | **0.5934** | 3,304/5,000 |

Both design metrics improve for all five backbones, while acceptance increases for HD, DiffPlanner, and GSDiff–SSIO, decreases for CE2EPlan, and changes little for ChatHD. Shared feedback therefore produces a consistent direction of design response, with backbone-dependent gains and measurement coverage. Equation (D.1) defines the normalized gains. The next case examines how completing a missing design relation affects the meaning of that response.

### E.6 Completing Design Relations: The GSDiff Case

#### E.6.1 From Native Structure to Access Relations

What happens when a generator's representation omits a relation required by the design theory? GSDiff (Hu et al., 2025a) learns RPLAN-derived corner and structural-edge graphs. Its preprocessing merges wall and door pixels into one structural mask; room cycles recover spatial partitions, and bubble-graph edges denote shared room boundaries. The resulting geometry specifies adjacency but leaves interior access undetermined. Other bindings also recover access in different ways: HD supplies generated door polygons, ChatHouseDiffusion supplies door-semantic pixels, and CE2EPlan and DiffPlanner use geometry-based inference (Appendix C.2). GSDiff–SSIO examines a learned completion route for this missing relation.

Within normalization $\mathcal{A}_m$, the frozen map $\mathcal{C}_\phi$ scores feasible room pairs from geometry, room types, and local context. Selected pairs receive geometry-anchored openings before construction of the common rectangular convex-atom access graph. Training uses 2,000 RPLAN plans, calibration uses 500, and fidelity evaluation uses a disjoint 500-plan split. Supervision consists of local access labels; SSIO subsequently computes integration and the terminal reward from the completed graph. Thus $\mathcal{C}_\phi$ learns the missing relation rather than predicting a design score. Both GSDiff–SSIO arms share this frozen map. SSPT-PPO updates only the main diffusion model, with the topology encoder and terminal edge stage fixed.

#### E.6.2 Access Recovery and Measurement Fidelity

How closely does completion reproduce the measured design organization? The held-out results meet three of nine prespecified criteria (Table 22). Access-pair recall reaches 0.9792 and precision is 0.8121; 97.2% of held-out plans have no opening-realization issues. Median absolute paired PIA error is 0.3454 and anchor-score correlation is 0.2726: recovering many access pairs still leaves substantial changes in the integration structure. The completed system's absolute scores consequently describe the combination of generated geometry and learned access completion, motivating its separate case-study role.

Table 22: **Access-completion fidelity on 500 held-out plans.** Thresholds were fixed before evaluation. Pair recovery, opening realization, and preservation of integration are tested separately.

| Criterion | Required | Observed | Met |
|---|---|---|---|
| Extra pairs introduced by opening realization | $= 0$ | 0 | Yes |
| Plans with opening-realization issues | $\leq 0.020$ | 0.028 | No |
| Access-pair precision | $\geq 0.900$ | 0.8121 | No |
| Access-pair recall | $\geq 0.900$ | 0.9792 | Yes |
| Access-pair F1 | $\geq 0.900$ | 0.8879 | No |
| Unambiguous-candidate recall | $\geq 0.995$ | 1.000 | Yes |
| Median absolute paired PIA error | $\leq 0.050$ | 0.3454 | No |
| $R_{\text{anchor}}$ Spearman correlation | $\geq 0.900$ | 0.2726 | No |
| Room-integration Spearman correlation | $\geq 0.900$ | 0.5405 | No |

#### E.6.3 Which Information Determines the Completed Scores?

We isolate how room-type information and connectivity decoding affect the inferred functional organization while keeping generated geometry fixed. A separate 1,000-condition experiment generates Base and +SSPT-PPO geometries once and rescores each through four frozen completion routes. All eight arm–route combinations accept 538 common conditions; Table 23 compares this shared subset.

We vary the information given to the frozen room-pair predictor, replacing excluded features with their training means so their standardized values become zero. The geometry/context route removes room-type and same-type features; the type route removes the remaining predictor features. All routes still use geometry to identify candidate room pairs and place openings. Separately, connected

decoding adds scored feasible pairs to connect components where possible, whereas raw decoding keeps only pairs selected by the predictor's threshold.

Table 23: **Access-source attribution on 538 common conditions from a separate 1,000-condition experiment.** Feature restrictions apply to the frozen pair predictor; every route retains geometric candidate and opening construction.

| | IPD ↓ | | median PIA ↑ | |
|---|---|---|---|---|
| Predictor features / decoding | Base | +SSPT-PPO | Base | +SSPT-PPO |
| Full / connected | 0.0362 | 0.0351 | 0.5882 | 0.5984 |
| Full / raw | 0.0275 | 0.0302 | 0.6144 | 0.6299 |
| Geometry/context / connected | 0.1102 | 0.1087 | 0.2781 | 0.2877 |
| Type / connected | 0.0407 | 0.0414 | 0.6093 | 0.6165 |

Removing type features lowers Base median PIA from 0.5882 to 0.2781, whereas the type route retains 0.6093. Room-type information therefore accounts for much of the strong Public anchoring assigned by completion, within the geometric candidate structure. Raw decoding gives higher median PIA than connected decoding in both arms, so the high anchoring is not explained by connectivity repair alone. The post-training IPD response also depends on the completion route: it improves with full connected and geometry/context connected scoring, and worsens with the other two routes on this common subset.

### E.6.4 Post-Training through the Fixed Completion Interface

The formal five-block experiment uses 5,000 conditions, separately from the attribution replay above. With the full frozen completion map, IPD changes from 0.0386 to 0.0360, median PIA from 0.5766 to 0.5934, and acceptance from 3,186 to 3,304 outputs (Table 21). Across the five blocks, Base IPD is $0.0386 \pm 0.0005$ and block median PIA is $0.5771 \pm 0.0047$, with population standard deviations.

Among the 2,462 conditions accepted by both arms, paired $\Delta PIA$ has mean $+0.0167$ and median zero: 1,225 improve, 248 tie, and 989 worsen. The overall gain thus does not extend to every matched plan. The generator learns from design feedback while the access-completion model stays frozen; Table 22 separately evaluates how well that model reconstructs real access organization. This case separates two requirements for architectural feedback: recovering access relations faithfully and using a fixed measurement to improve the generator.

# F GENERATIVE AI FOR DESIGN: LITERATURE REVIEW AND SYNTHESIS

To place the floorplan study in a broader design context, we examine how other tasks connect domain knowledge to generative modeling. Our review of 55 studies across 14 design categories follows design purposes from their knowledge sources to their computational and learning roles through three review questions:

1. **RQ1: What counts as design progress, and what evidence supports it?** We compare synthesis-oriented and purpose-oriented contributions across research communities (Section F.2).
2. **RQ2: Where do design criteria come from, and how is their meaning made computable?** We trace knowledge sources to analytical objects and measurements (Section F.3).
3. **RQ3: How do design criteria change generation, and what makes them reusable learning objectives?** We distinguish assessment, inference, and model updates, then examine adaptation across generators (Section F.4).

Together, these questions identify where design-theory translation is established, where its connections remain incomplete, and how the floorplan study addresses this broader problem (Section F.5). Appendix G develops the construction principles and future directions that follow.

## F.1 REVIEW SCOPE AND EVIDENCE CODING

**Selection and coverage.** We used purposive sampling across work published or released from 2021 to September 13, 2026, starting from recent systems and tracing representative synthesis, evaluation, control, and learning approaches. All 55 studies in 14 design categories enter the full-sample summaries (Figures 13 and 16). Fifty have method-level evidence and detailed mapping coordinates; three abstract-level records and two contextual reviews or benchmarks retain their distinct evidence status. Earlier design theories and general learning methods supply context outside this inventory. Counts describe this selected sample; comparisons concern mechanisms and evidence.

**What we record.** Each record identifies the year, publication setting, design category, contribution type, represented objects, knowledge source, criterion use, updated component, and validation evidence. We distinguish generator development, evaluation, and human–AI workflows by their actual methods. Publication setting locates a research tradition; the stated objective and experiments establish what the work contributes. Section F.6.2 gives the coding key alongside the complete inventory.

**Computational roles and knowledge sources.** The maps distinguish implicit representations, explicit elements, and explicit relations, then trace criteria through example fitting, evaluation, inference, and learning. Learning includes supervision, losses, and rewards, with the updated component recorded separately. These are computational roles, not theory or quality rankings. RQ2 independently distinguishes domain theories, established principles, author-defined criteria, task specifications, physical models, and preference judgments.

**Full-sample classification.** Research emphasis distinguishes synthesis quality and specified-condition validity, domain-purpose performance, and contributions examining both. It describes what a study investigates or evaluates; the criterion's destination records what it actually changes. A study may assess a design purpose while its generator still learns only from examples. We identify the knowledge source of each criterion and its documented uses, including feedback mediated by learned evaluators or surrogates. Example-based records identify fitting without a separately established design criterion; unresolved provenance remains explicit. Expert validation is kept separate from training-label provenance. Sources and destinations can overlap.

**Evidence and interpretation.** Coding follows documented representations, losses, reward definitions, update rules, and evaluation protocols. Fixed conditioning, interpretation of underspecified requests, interactive revision, and preference learning are recorded separately. Reported aims are distinguished from demonstrated computation and learning. Coverage is task-specific: editorial examples concern single pages, visualization examples principally charts, and exhibition examples spatial

layouts. Figures summarize the comparisons; Section F.6.2 preserves all study names, citations, and evidence status.

### F.2 RQ1: Design Progress and Its Evidence

**Two standards of progress.** Generative modeling asks whether an artifact is faithful and controllable; design asks what it accomplishes. Visual and geometric quality make a proposal recognizable, coherent, or constructible. Design quality also concerns how its elements serve a purpose: communicating an argument, supporting movement, assembling a garment, or carrying a load. Statistical similarity and functional performance therefore require different evidence (Regenwetter et al., 2023). Our question is what should guide learning once a generator produces viable objects.

**Different research starting points.** A recurring contrast in the surveyed work is what defines progress. AI synthesis papers foreground representations, conditioning, and generation quality: LayoutTransformer models page layouts, DeepCAD construction sequences, and HouseDiffusion vector coordinates (Gupta et al., 2021; Wu et al., 2021; Shabani et al., 2023). Domain-oriented studies foreground use, professional judgment, or physical performance (Regenwetter et al., 2023; Park et al., 2024). The former can advance generation without explicitly measuring the design purpose; the latter can articulate that purpose without making it a generator's learning objective.

**Model advances and design evidence.** Museum-CGAN illustrates the separation: a conventional conditional GAN learns adversarial and perceptual objectives, while questionnaires assess the resulting museum proposals (Tang et al., 2024a). TopoDiff, published at an AI venue, brings mechanical performance into diffusion sampling (Mazé & Ahmed, 2023); interdisciplinary DRL-Urban directly trains a graph policy for planning goals (Zheng et al., 2023). Publication community thus identifies a research setting, while the objective and validation protocol establish the actual contribution. Figure 14 makes these distinctions visible; Table 25 separates data volume, design cases, and human assessments. The opportunity is to connect domain-grounded purposes to reusable model improvement, bringing together the strengths of both research traditions.

Table 24: Synthesis quality and design purpose ask different questions about the same artifact.

| Design object | Synthesis-oriented question | Purpose-oriented question |
|---|---|---|
| Graphic layout | Are elements and placements plausible? | Does their organization establish a readable hierarchy? |
| Mechanical part | Is the geometry valid and similar to examples? | Does the structure perform under its intended loads? |
| Garment | Can a coherent pattern be generated? | Do panel and sewing relations support assembly and wearing? |
| Floorplan | Are rooms and boundaries geometrically coherent? | Does their access organization support the intended use of spaces? |

**Comparing purposes across tasks.** We next examine what each design organizes and what that organization accomplishes. This makes the difference between synthesis and design progress concrete, while preserving the contributions of systems that already optimize a domain purpose.

#### F.2.1 Organization for Communication and Interaction

Graphic and information designs organize elements to direct attention, convey meaning, and support action. A coherent arrangement is the starting point; reading hierarchy, scientific fidelity, and interaction introduce more specific standards of success.

**Posters and advertising.** PosterLLaVa learns placements from reference layouts; LaySPA post-trains a generator with explicit spacing, alignment, overlap, and text–underlay rewards; PosterReward adapts an image generator through learned judgments of rendered posters (Yang et al., 2024b; Li et al., 2026b; Lai et al., 2026). The alternatives are substantively different: imitate arrangements, optimize declared relations, or learn a holistic judgment.

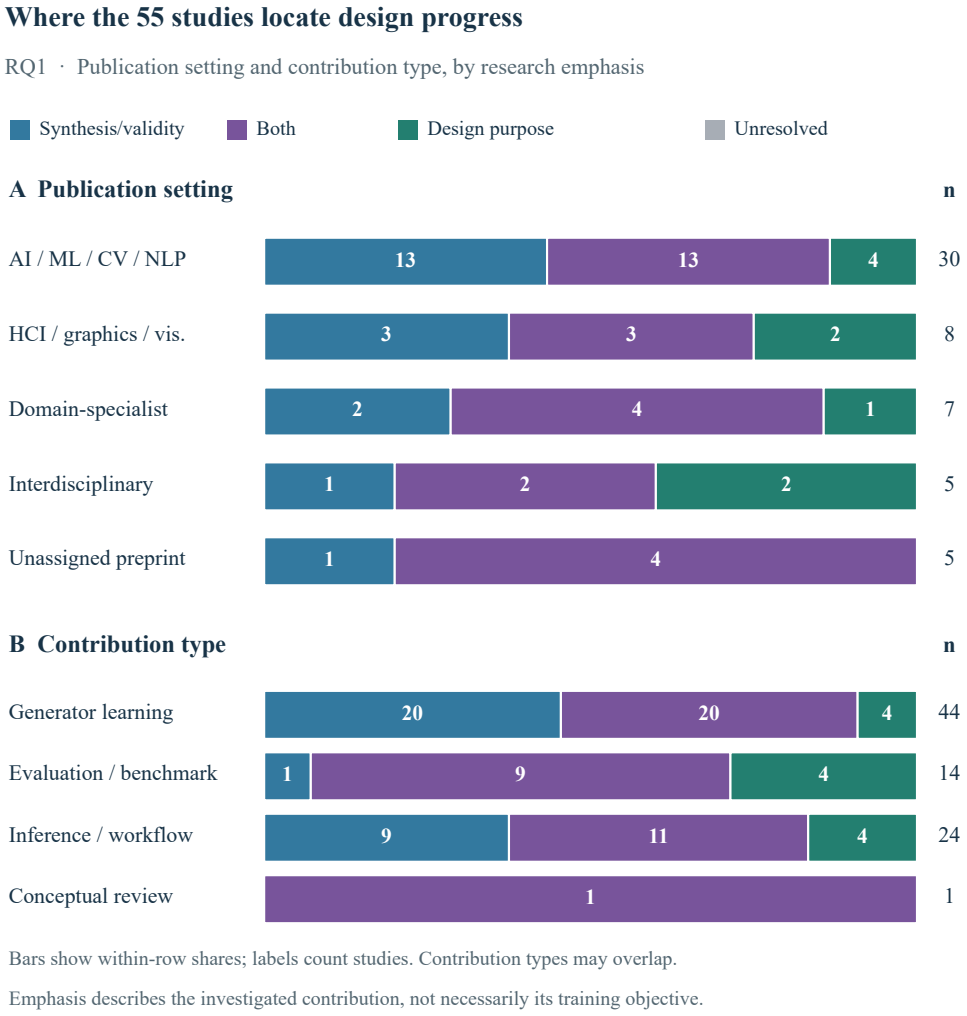



Figure 13: **Research emphasis across all 55 reviewed studies.** Publication setting and contribution type are classified independently of the investigated objective. A study may contribute to multiple types in panel B. Study-level assignments and references appear in Section F.6.2; Figure 14 explains representative mechanisms.

**Editorial and document layout.** LayoutTransformer models page-element sequences; LayoutDM adds relation energies during sampling, LayoutFormer++ constrains decoding, and LayoutNUWA generates editable code (Gupta et al., 2021; Inoue et al., 2023; Jiang et al., 2023; Tang et al., 2024b). These mechanisms control placement and expose elements needed to study information hierarchy, but their reported objectives do not directly measure reading or argumentation.

**Scientific diagrams.** AutomaTikZ and DiagramAgent improve graphical programs through compilation or checking; AutoFigure critiques and refines an entity–relation blueprint before rendering (Belouadi et al., 2024; Wei et al., 2025; Zhu et al., 2026). The inspectable blueprint separates scientific organization from appearance, although correction chiefly occurs during inference.

**Data visualization.** LIDA translates analytical goals into executable charts; ChartGPT produces data transformations and chart specifications; Data Formulator 2 exposes transformations and visual encodings for revision (Dibia, 2023; Tian et al., 2025; Wang et al., 2025). Here organization serves a data question, and users revise analytical meaning together with visual form.

**Research setting, computational contribution, and evidence**

RQ1 · What counts as progress, and what establishes its design value?

| Study / research setting | Computational intervention | Validation target |
|---|---|---|
| **HouseDiffusion**<br>CVPR 2023 · AI | Denoising and coordinate fitting | Realism and graph compatibility |
| **TopoDiff**<br>AAAI 2023 · AI | Mechanics-guided diffusion sampling | Compliance and material feasibility |
| **UIClip**<br>UIST 2024 · HCI | Principle- and preference-trained evaluator | Agreement with professional judgments |
| **DRL-Urban**<br>Nat. Comput. Sci. 2023 · IX | Purpose-driven planning-policy training | Planning measures and professional comparisons |
| **Museum-CGAN**<br>PLOS ONE 2024 · IX | Adversarial and perceptual fitting | Generated proposals and 264 questionnaire responses |
| **Small-residence analysis**<br>J. Build. Eng. 2026 · Domain | Layout generation and design assessment | Dataset analysis and one residential case |

**Objectives and evidence reveal what each research setting contributes.**

IX: interdisciplinary. Colors identify design families; named studies illustrate distinct routes.

Figure 14: **Research setting, computational intervention, and validation.** Representative cases explain the 55-study distribution in Figure 13; the small-residence analysis adds architectural context outside that coded sample. References and validation details appear in Section F.6.2 and Table 25.

**Interfaces and interaction.** UIClip uses design principles and professional preferences to train a screenshot evaluator; UICoder fine-tunes a UI-code generator through compilation and prompt-relevance feedback (Wu et al., 2024a;b). One learns judgment and the other changes production, while static presentation, executable implementation, and successful interaction remain distinct validation targets.

**Packaging.** A kumis-packaging workflow combines cultural elements, shape grammar, requirement weighting, prompting, and proposal selection; a craft-beer study compares visual and brand expression across tools (Zhao et al., 2026; Chen & Jang, 2026). Both center cultural communication and identity in evaluating proposals, without updating the underlying generator or addressing structural packaging engineering.

### F.2.2 Organization for Construction and Physical Performance

For physical designs, organization must support assembly and behavior. Valid geometry establishes an object that can be manipulated or constructed; performance criteria determine whether that object is useful under its intended conditions.

**Industrial products and CAD.** DeepCAD learns construction sequences, BrepGen boundary geometry and associations, and Text2CAD language-conditioned parametric construction (Wu et al., 2021; Xu et al., 2024; Khan et al., 2024). MO-PaDGAN instead trains with performance and diversity, while TopoDiff guides sampling through mechanical criteria (Chen & Ahmed, 2021; Mazé & Ahmed,

2023). Because similarity does not establish engineering value (Regenwetter et al., 2023), both the measured property and its effect on generation matter.

**Garments.** DressCode, SewingLDM, and AIpparel represent panels and sewing relations—structure tied to assembly, not only appearance (He et al., 2024; Liu et al., 2025b; Nakayama et al., 2025). Reference patterns teach structured generation, and SewingLDM adds integrity terms to its representation model. Simulation, reconstruction, and user studies then test different aspects of assembly and wearing.

**Textiles and patterns.** Textile diffusion and Knit-FLUX target pattern appearance and knitted-fabric imagery, whereas functional-textile work represents interlacing and yarn fields under surrogate performance losses (Karagoz et al., 2023; Liu et al., 2025c; Tang et al., 2026). This separates organizing a surface motif from organizing material for physical behavior; implementation and validation evidence are recorded in the inventory.

#### F.2.3 Organization for Spatial and Sequential Experience

Spatial and temporal designs organize routes through space or events. Accessibility, playability, continuity, and narrative transition supply purposes that visual consistency alone cannot establish.

**Urban and landscape design.** UrbanGAN fits block morphology and evaluates spatial attributes; DRL-Urban trains policies for accessibility, ecology, and road efficiency (Fedorova, 2021; Zheng et al., 2023). PlantoGraphy connects intentions, scene graphs, layouts, and rendering interactively, whereas GenFusion combines diffusion candidates with performance prediction and evolutionary search (Huang et al., 2024; Gao et al., 2025). Design purpose thus enters through policy learning, interaction, search, or assessment.

**Film, storyboards, and animation.** TaleCrafter controls character layouts; CineVerse plans scenes and shots for consistent keyframes; ShotPlan learns shot-control tokens; and STAGE preference-optimizes frame pairs before a separate video model renders transitions (Gong et al., 2023; Phung et al., 2026; Guo et al., 2026; Zhang et al., 2026b). Progress in identity consistency, shot organization, and transitions therefore requires evidence at different levels. That evidence must also distinguish what planning or control achieves from what a model learns through updates.

**Game levels.** Controllable PCGRL trains from path, region, and playability objectives; later work scales this setting and adds language control (Earle et al., 2021; 2024; Baek et al., 2025). DiffusionCraft instead fits local patterns from one level and evaluates playability afterward (Dai et al., 2024). Purpose-driven RL and example-fitting diffusion therefore coexist; the distinction is which design capability enters learning.

**Exhibition spaces.** Museum-CGAN fits layouts adversarially and perceptually before human evaluation; Rule-Exhibit turns connectivity, salience, and artwork importance into topology and placement supervision; later work compares image and spatial properties across generators (Tang et al., 2024a; Xu et al., 2025; Li et al., 2026c). These are different interventions: judging proposals, constructing design-informed targets, and comparing systems. Rules can therefore enter learning without an RL reward.

**RQ1 finding: design interest and design learning must be distinguished.** Of the 55 records, 20 emphasize synthesis or validity, nine design purpose, and 26 examine both (Figure 13). Both emphases occur across publication communities. Their presence in a study, however, does not establish a purpose-directed training objective: professional assessments may follow example fitting, and performance criteria may guide inference without changing model weights. Thus, the revealing comparison is how design value becomes an evidenced model capability. Figure 15 exposes the represented objects and computational roles behind that connection.

### Representing relations and using design criteria

Named examples across design domains, 2021-2026

| Study (year) | How relations are represented | | | Criterion used in |
|---|---|---|---|---|
| | Implicit | Elements | Relations | |
| **Graphic, information and interface design** | | | | |
| LayoutTransformer (2021) | ○ | ● | ○ | Fitting |
| LaySPA (2026) | ○ | ○ | ● | Learning |
| Data Formulator 2 (2025) | ○ | ○ | ● | Inference |
| AutoFigure (2026) | ○ | ○ | ● | Inference |
| UIClip (2024) | ● | ○ | ○ | Learning |
| **Products, garments and materials** | | | | |
| Text2CAD (2024) | ○ | ○ | ● | Evaluation |
| MO-PaDGAN (2021) | ○ | ● | ○ | Learning |
| DressCode (2024) | ○ | ○ | ● | Evaluation |
| SewingLDM (2025) | ○ | ○ | ● | Learning |
| TextileDiffusion (2023) | ● | ○ | ○ | Fitting |
| KumisPackaging (2026) | ○ | ● | ○ | Inference |
| **Spatial and temporal design** | | | | |
| DRL-Urban (2023) | ○ | ○ | ● | Learning |
| PlantoGraphy (2024) | ○ | ○ | ● | Inference |
| STAGE (2026) | ○ | ○ | ● | Learning |
| DiffusionCraft (2024) | ○ | ● | ○ | Evaluation |
| IPCGRL (2025) | ○ | ○ | ● | Learning |
| Rule-Exhibit (2025) | ○ | ○ | ● | Learning |
| **Floorplan generation** | | | | |
| HouseDiffusion (2023) | ○ | ○ | ● | Evaluation |
| ChatHouseDiffusion (2024) | ○ | ○ | ● | Inference |
| GSDiff (2025) | ○ | ○ | ● | Learning |
| FloorPlan-LLaMa (2025) | ● | ○ | ○ | Learning |
| Floorplan RLVR (2026) | ○ | ○ | ● | Learning |
| GreenPlanner (2026) | ○ | ○ | ● | Learning |
| **Professional image evaluation** | | | | |
| Q-Judger (2026) | ● | ○ | ○ | Learning |

Elements: explicit parts / attributes. Relations: links, programs or constraints.

Inference includes selection and repair; Learning includes non-generator models.

Relations may appear in model inputs, intermediate states, or outputs.

Figure 15: **Representation and criterion use are distinct.** Named studies connect represented organization to criterion use. Figure 17 distinguishes theory-derived measures from other criteria; Section F.6.2 gives the complete inventory.

### F.3 RQ2: From Knowledge Sources to Design Measurement

**Evaluate the purpose of the particular design.** IDEA-Bench tests professional task completion; Qwen-Image-Bench and ArtiMuse train evaluators from professional rubrics or aesthetic annotations; PosterReward targets poster quality (Liang et al., 2025; Li et al., 2026a; Cao et al., 2026; Lai et al., 2026). The relevant purposes differ: a poster organizes attention and communication, while a dwelling organizes access, shared activities, and privacy. Appearance and organization contribute according to the task; neither a universal aesthetic score nor geometric validity defines success across domains.

**Identify where the criterion comes from.** A design-related score need not derive from domain design theory. We reserve *theory-derived* for a traceable connection between an established theory, its analytical objects, and the interpreted property being measured. Design principles offer guidance that authors operationalize; task specifications define requirements; physical models predict behavior; and preference labels teach a judgment. These sources can coexist in one system. Figure 16 classifies all 55 records by knowledge source and computational destination; Figure 17 explains representative connections.

**Separate theoretical meaning from implementation.** UIClip uses contrast, repetition, alignment, and proximity (CRAP) principles to construct perturbations, then learns judgment from screenshots and preferences. LaySPA explicitly computes authored spacing, alignment, overlap, and text–underlay rewards (Wu et al., 2024a; Li et al., 2026b). TopoDiff uses structural mechanics; DRL-Urban operationalizes planning concepts and service standards, including the 15-minute life circle (Mazé & Ahmed, 2023; Zheng et al., 2023). Space syntax supplies both an analytical representation and an interpretation of integration before either is embedded in a learning system (Hillier & Hanson, 1984; Hillier, 1996).

**Architectural assessment and its remaining gap.** Floorplan evaluation covers image similarity, geometry, program constraints, and architect-informed rules (Nauata et al., 2021; Shabani et al., 2023; Zeng et al., 2025a; Luo et al., 2024; van Engelenburg et al., 2025; Park et al., 2024). FloorPlan-LLaMa learns functionality, flow, and overall judgments from expert labels (Yin et al., 2025), but the spatial basis of each score remains implicit. The evaluator ranks perceived quality without exposing the access organization responsible for it.

**Space syntax connects organization to use.** Space syntax analyzes spatial configuration through explicit relations and measures (Hillier & Hanson, 1984; Hillier, 1996; Jiang et al., 2000), grounded in architectural studies of domestic organization and design transformation (Hanson, 1998; Edgü & Ünlü, 2003; Helme et al., 2014; Eloy & Guerreiro, 2016; Azizi et al., 2023). SSIO recovers convex spaces and access relations, then measures relative accessibility. Theory defines the property; residential corpora provide the empirical reference needed to interpret it.

Residential evidence also gives these properties context. Studies examine living-room-centered configurations in Korean apartments (Byun & Choi, 2016), collective activities in integrated spaces in Inuit households (Dawson, 2003), and private and shared experiences of living rooms (Rechavi, 2009). Together, they motivate interpreting accessibility through domestic use and cultural context, which our cross-corpus analysis makes explicit in the reference distributions.

**RQ2 finding: explicit computation and theoretical grounding are independent.** A graph exposes relations, a rule supplies a computable score, and preference labels supply judgments. Interpreting a score still requires an account of what it measures and why that property matters for the design task. Theory-derived measures obtain this account from the theory's analytical objects and interpretation. Figure 18 locates computational interfaces and signal use; Figure 17 distinguishes the knowledge behind them. The research need therefore depends on what is missing: a computational form for existing knowledge, a learning method for an established measure, or evidence that a new measure captures the intended design purpose.

**Knowledge sources behind computable criteria.** Specified requirements occur in 41 records, authored criteria in 18, and ratings or preferences in 17; these sources overlap (Figure 16). Established principles and physical models also enter learning. The named space-syntax use in this inventory concerns museum assessment and is documented at abstract level (Li et al., 2026c).

**From knowledge sources to computational use: all 55 studies**

RQ2–RQ3 · A source can support assessment, inference, or model learning

| | Example fitting | Assessment | Inference | Auxiliary learning | Generator training | Generator post-training | n |
|---|---|---|---|---|---|---|---|
| Domain theory | | 1† | | | | | 1 |
| Established principles | | 3 | 3 | 2 | 1 | 1 | 4 |
| Authored criteria | | 14† | 5 | 3 | 5 | 3 | 18 |
| Specified requirements | | 28† | 24† | 5 | 7 | 4 | 41 |
| Physical models | | 6† | 4† | 2 | 2 | 1 | 6 |
| Ratings / preferences | | 15† | 2 | 6 | | 3 | 17 |
| Example-based only | 3 | | | | | | 3 |
| Unresolved source | | 1† | | | | | 1 |

Cells count documented links; n counts unique studies with each source.

† Includes abstract-level or contextual evidence; these records remain identifiable.

**Knowledge origin and learning destination are separate dimensions.**

Figure 16: **Knowledge provenance and computational destination across all 55 studies.** Cells count documented source-to-use links, including learned implementations; both dimensions may overlap. Marked cells include abstract-level or contextual evidence. Our framework is excluded from these counts and appears separately in Figure 17.

### F.4 RQ3: From Design Criteria to Reusable Generator Learning

The preceding analysis identifies what makes a design successful. We now trace that purpose into model improvement by distinguishing the objective, training stage, and component updated.

#### F.4.1 Training Stage and Design Objective

Pretraining, continued pretraining (CPT), supervised fine-tuning (SFT), preference optimization, and RL describe how or when a model learns, not what design property it learns. CPT may extend domain modeling (Gururangan et al., 2020); SFT may itself be post-training; RL may train a policy from scratch. UICoder combines compiler- and relevance-filtered SFT with preference optimization (Wu et al., 2024b). We therefore record the optimized property and updated component alongside the training route.

**From knowledge provenance to a computational design criterion**

RQ2 · What gives a criterion its design meaning?

| Study | Knowledge source | → | Analytical form | → | Used in |
|---|---|---|---|---|---|
| UIClip | Design principles and preferences<br>CRAP; designer ratings | | Screenshot features; learned quality score | | Evaluator learning |
| LaySPA | Authored layout criteria<br>Spacing, alignment, pairing | | Element placements; explicit layout rewards | | LLM post-training |
| TopoDiff | Structural mechanics<br>Compliance; feasibility | | Material fields; performance predictors | | Diffusion sampling |
| DRL-Urban | Planning concepts and standards<br>Service, ecology, traffic | | Planning graph; computable use criteria | | Planning-policy training |
| FloorPlan-LLaMa | Architect preferences<br>Functionality, flow, quality | | Image features; learned preference score | | Autoregressive post-training |
| Floorplan RLVR | Specified requirements<br>Adjacency, area, validity | | Polygons and graph; requirement checks | | Autoregressive post-training |
| Our framework | Space-syntax theory<br>Real-design references | | Recovered access IR; measured integration | | Heterogeneous diffusion post-training |

**Explicit computation and theory-derived meaning are different properties.**

IR: intermediate representation. Sources may coexist; rows identify the criterion discussed.

Figure 17: **Representative knowledge-to-computation connections.** These cases explain the categories in Figure 16. Our pathway carries space-syntax objects and measurements through empirical grounding into heterogeneous-generator post-training. Study references are given in the accompanying analysis and inventory.

LayoutTransformer fits structured layouts and HouseDiffusion denoising targets (Gupta et al., 2021; Shabani et al., 2023); their examples carry organization even when the objective does not measure its purpose. LayoutNUWA and PosterLLaVa similarly fine-tune pretrained models on reference specifications (Tang et al., 2024b; Yang et al., 2024b). Explicit design purpose can also enter initial training: MO-PaDGAN adds performance and diversity objectives, DRL-Urban rewards accessibility, ecology, and road efficiency, and Rule-Exhibit converts design rules into topology and placement supervision (Chen & Ahmed, 2021; Zheng et al., 2023; Xu et al., 2025). The distinction lies in the property driving updates, whether supplied through examples, losses, rewards, or constructed targets.

### F.4.2 Where Design Feedback Acts

A criterion may evaluate outputs, guide sampling, select candidates, train an evaluator, optimize prompts, or update the generator. LayoutDM and TopoDiff act during inference; UIClip trains an evaluator; GDB-Reward a prompt policy; LaySPA and PosterReward the generator (Inoue et al., 2023; Mazé & Ahmed, 2023; Wu et al., 2024a; Deganutti et al., 2026; Li et al., 2026b; Lai et al., 2026). Figure 20 distinguishes these destinations: better evaluation or selection does not itself change the generator.

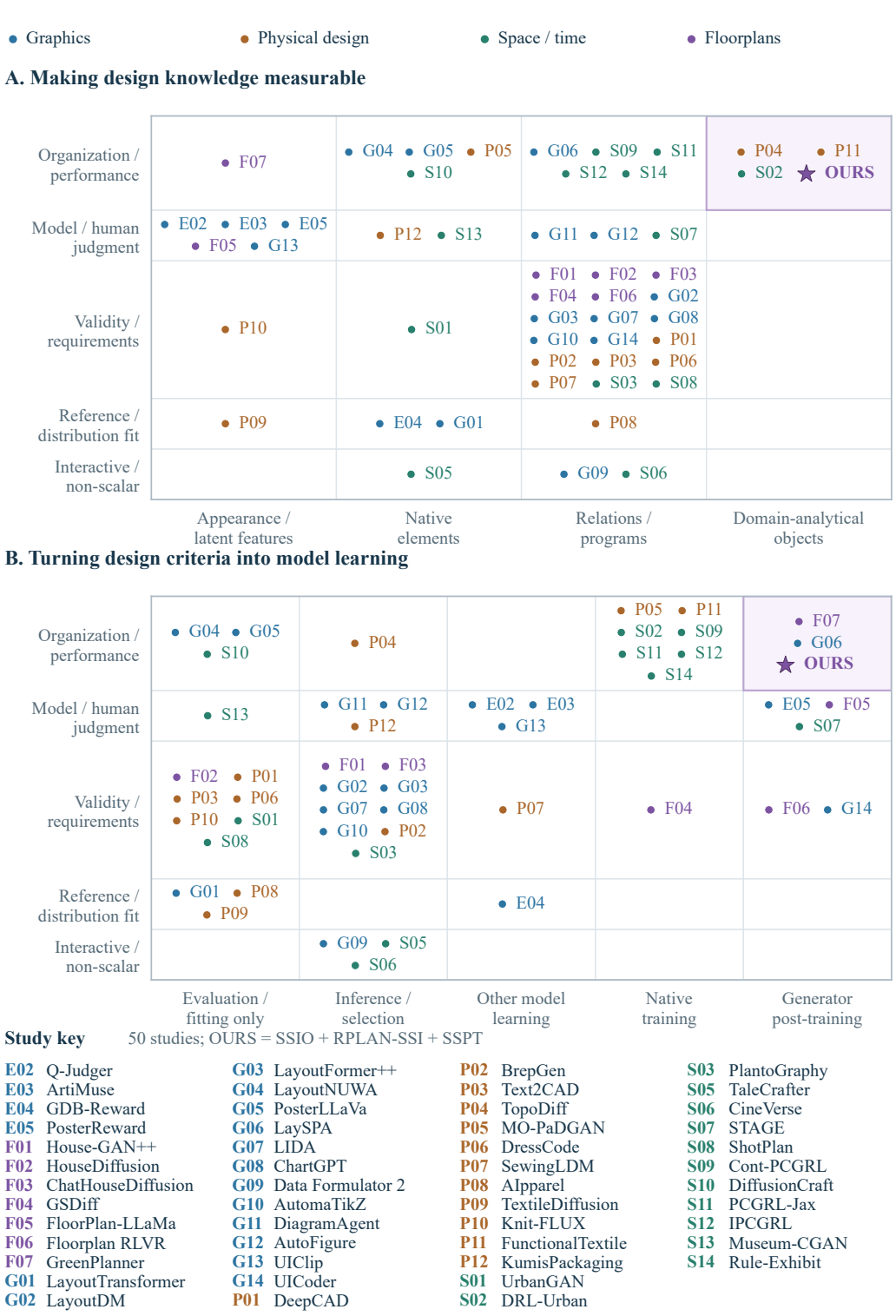



Figure 18: **From design signals to model learning.** Fifty studies are classified by (A) analysis interface and signal basis and (B) the component affected. Organization/performance includes authored criteria and physical objectives as well as theory-derived measures; Figure 17 separates these foundations. The star denotes our SSIO/RPLAN-SSI/SSPT pathway. The key names each study; Section F.6.2 provides references.

Differentiable rewards, policy gradients, and preference or group-relative objectives provide adaptation mechanisms (Clark et al., 2024; Prabhudesai et al., 2023; Fan et al., 2023; Black et al., 2024; Zhao et al., 2024; 2025; Hu et al., 2025b; Christiano et al., 2017; Ouyang et al., 2022; Wallace et al., 2024; Yang et al., 2024a; Liu et al., 2025a); DPPO extends diffusion policy gradients to control and robotics (Ren et al., 2025). Reward-ranked and synthetic-data retraining offer further routes (Dong et al., 2023; Ferbach et al., 2024; Alemohammad et al., 2024). These methods specify how a signal changes a model. Design theory supplies the property to improve and the objects on which it is measured.

#### F.4.3 Post-Training Existing Generators

Purpose-driven initial training builds a generator around a design objective. Post-training asks how that objective can improve an established generator while retaining its geometric quality and response to conditioning.

In architecture, prior work applies visibility or space-syntax analysis to reconstructed plans, small residences, and museum layouts (Heo & Cho, 2023; Man et al., 2026; Li et al., 2026c). FloorPlan-LLaMa instead post-trains an autoregressive image-token generator through learned architect preferences; floorplan RLVR adapts an autoregressive layout model through graph, area, and validity rewards (Yin et al., 2025; Lara et al., 2026). The former leaves the spatial basis implicit; the latter verifies specified conditions rather than a theory-defined design property. GreenPlanner targets energy and feasibility in a flow generator (Zeng et al., 2026); other methods alter conditions, outputs, topology, or latent control (Zeng et al., 2025b; Zong et al., 2026; Zhang et al., 2026a; Güldilek et al., 2026). SSPT connects explicit architectural measurement to post-training across heterogeneous floorplan diffusion policies.

PosterReward adapts several image-generator backbones, and UICoder several UI-code backbones (Lai et al., 2026; Wu et al., 2024b). Our setting additionally requires recovering spaces and access relations from different floorplan formats, so that one accessibility measure can guide generators with different policies (Figure 19). Each generator retains its own synthesis and update mechanisms.

**RQ3 finding: the learning destination determines the capability changed.** Assessment and selection can improve delivered artifacts; generator updates change the model producing them. Figure 16 traces each criterion to this destination. Reusing a design objective across generators requires both a common measure of that objective and an update method for each generator.

### F.5 Cross-RQ Synthesis: The Floorplan Study in Context

The three findings expose a connected research problem: a design purpose needs evidence of value (RQ1), a computational criterion needs a traceable domain meaning (RQ2), and that criterion needs a learning interface to improve established generators (RQ3). Floorplan generation brings these connections into one experimentally accessible setting.

**Existing generation capabilities make the question testable.** Floorplan methods condition on boundaries, room programs, relations, partial specifications, and activities (Wu et al., 2019; He et al., 2022; Hu et al., 2020; Nauata et al., 2021; Para et al., 2021; Zhang et al., 2024; Hong et al., 2024; Wang et al., 2023). Explicit room and wall modeling (Sun et al., 2022; Liu et al., 2022; Tang et al., 2023; Wang & Pajarola, 2025; Hu et al., 2025a) supports heterogeneous diffusion, autoregressive, and multimodal generators (Shabani et al., 2023; Zeng et al., 2024; Wang & Pajarola, 2026; Lian et al., 2026; Qin et al., 2024; Shiohara & Yamasaki, 2026; Qin et al., 2026). This diversity motivates studying whether architectural feedback can improve established generators without replacing their architectures; our experiments test this question on five diffusion backbones.

Graph-constrained generation (Upadhyay et al., 2022; Aalaei et al., 2023) and neural-guided optimization (Zheng & Petzold, 2023) exemplify different routes to satisfying specified relations. A shared theory-defined evaluation can assess the spatial organization produced by either route.

**The design objective is ready for explicit translation.** Space syntax supplies analytical objects and interpretable accessibility measures; residential corpora reveal their variation with use and context.

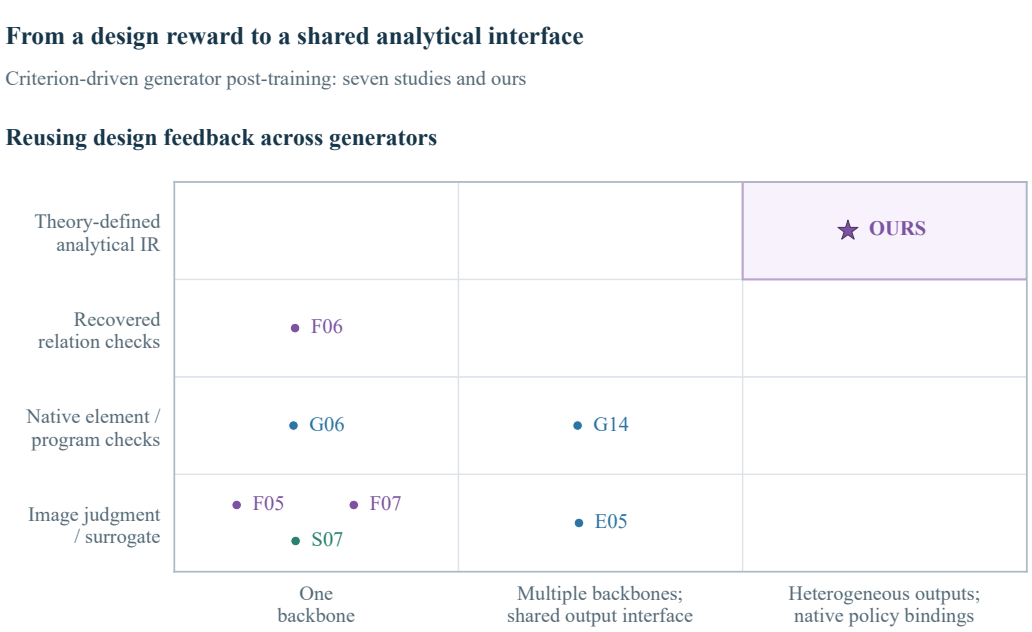



Figure 19: **Post-training scope and feedback interface.** Seven prior methods and our pathway are classified by their feedback basis and demonstrated generator adaptation. IR denotes intermediate representation. SSIO and SSPT connect theory-defined analytical objects to native policy updates across heterogeneous outputs.

Together with established heterogeneous generators, these resources make the theory-to-learning gap identified in Section F.4 experimentally tractable. Floorplan generation is therefore a timely test of design-theory translation: can shared architectural feedback improve existing generators toward the task's original purpose while retaining their required output quality?

**Implications and open directions.** The opportunity is to translate domain knowledge through the whole representation–evaluation–learning chain while retaining the generative capabilities already established. Where an appropriate theory exists, the next step is computational translation and empirical grounding; where it does not, constructing a useful representation and score must be accompanied by evidence that they capture the intended design purpose. Appendix G develops both the reusable construction procedure and this longer-term direction toward AI-native design theories (Section G.1).

### Which design objective changes which model?

A criterion entering learning does not identify the model being improved.

| Study / venue | Criterion | Updated component |
|---|---|---|
| **Design criteria in native model training** | | |
| MO-PaDGAN (2021)<br>Applied Soft Computing | Performance and diversity | GAN generator |
| DRL-Urban (2023)<br>Nature Computational Science | Service access, ecology and road efficiency | Planning policy trained from scratch |
| **Design criteria in representation or evaluator learning** | | |
| SewingLDM (2025)<br>ICCV | Panel and stitch integrity | Sewing-pattern representation model |
| UIClip (2024)<br>UIST | Contrast, repetition, alignment, proximity | Interface evaluator |
| **Design criteria in prompt-policy learning** | | |
| GDB-Reward (2026)<br>Preprint | Text, color and layout fidelity | Prompt policy; image generator frozen |
| **Design criteria in generator post-training** | | |
| LaySPA (2026)<br>ACL Industry | Spacing, alignment, text-underlay pairing | LLM layout generator |
| STAGE (2026)<br>CVPR | Cinematic transition preferences | Frame-pair predictor; video generator frozen |
| FloorPlan-LLaMa (2025)<br>ACL | Expert functionality, flow and quality | Autoregressive floorplan generator |
| Floorplan RLVR (2026)<br>ACL Findings | Graph and area requirements | Autoregressive floorplan generator |
| GreenPlanner (2026)<br>CVPR Findings | Energy and feasibility | Normalizing-flow floorplan generator |
| **SSPT (this work)** | Space-syntax integration | **Heterogeneous diffusion floorplan generators** |

Each row links a stated criterion to the component it updates.

Publication venues are shown to connect methodological and domain perspectives.

Figure 20: **A design criterion can update different components.** The examples separate native training, representation or evaluator learning, prompt-policy learning, and generator post-training. SSPT connects one architectural measurement to heterogeneous diffusion policies.

## F.6 Validation Evidence and Complete Study Inventory

### F.6.1 Validation at the Level of the Claimed Contribution

Different research priorities require different evidence. Held-out synthesis tests establish model behavior across examples; professional studies may examine decisions, experience, or use in a particular setting. Both are informative when matched to the claim. Table 25 separates training or source-data scale, independent test cases, and human evaluation, so that a large count in one category is not mistaken for evidence in another. We use documented losses, reward definitions, update rules, and evaluation protocols to distinguish reported design aims from demonstrated generator learning.

Table 25: Validation protocols answer different questions. Source-data volume, independent design cases, and human assessments are reported separately.

| Study / community | Reported scale or protocol | What the evidence tests |
|---|---|---|
| DRL-Urban / interdisciplinary (Zheng et al., 2023) | One synthetic scenario and two real communities; repeated policy runs; eight professional-planner comparisons; 100 design-trained evaluators. | Use-related planning objectives and professional comparisons across the evaluated settings. |
| Museum-CGAN / interdisciplinary (Tang et al., 2024a) | 100 images and 264 valid questionnaire responses; no conventional train/test split reported. | Perceived design outputs; questionnaire scale does not establish held-out generator generalization. |
| Small-residence analysis / specialist journal (Man et al., 2026) | 14,373 extracted small-layout samples; evaluation demonstrated on a 30 $m^2$ residential unit. | Dataset analysis and a design-assessment case, rather than large-scale model adaptation. |
| Rule-Exhibit / graphics (Xu et al., 2025) | Rule-generated supervision and quantitative comparisons; dataset size and split not established in the inspected report. | A rule-to-learning mechanism; the available report leaves independent testing breadth unspecified. |
| Data Formulator 2 / HCI (Wang et al., 2025) | Eight participants in a user study. | Iterative analytical workflow and interaction. |
| LaySPA / AI (Li et al., 2026b) | 3,000 training layouts per dataset; 1,000 and 905 test layouts. | Generator post-training and held-out layout criteria. |
| Kumis packaging / interdisciplinary (Zhao et al., 2026) | 106 cultural items, 38 selected source items, and three final packaging designs. | Knowledge-guided proposal development in a human–AI workflow. |

### F.6.2 Cited Study Inventory

Each entry identifies the work by name and citation, its represented objects and criterion, and the computational step affected by that criterion. IDs connect the figures to the corresponding studies. Abstract-level and contextual entries retain their evidence status. The added codes record research focus (F) and knowledge basis (K), independently of the original representation and criterion-use codes.

**Publication setting:** AI = AI/ML/CV/NLP; HG = HCI, graphics, and visualization; DS = domain-specialist; IX = interdisciplinary; PP = preprint without an assigned venue.

**Representation (R):** R0 = implicit in images or latent tokens; R1 = explicit elements and attributes; R2 = explicit relations, programs, or operational constraints. **Criterion use (P):** P0 = example fitting; P1 = separate evaluation; P2 = inference or selection; P3 = criterion-driven learning.

**Research focus (F):** S = synthesis/validity, D = design purpose, B = both, U = unresolved. **K:** T = domain theory, P = established principles, A = authored criteria, R = specified requirements, M = physical models, J = ratings/preferences, E = example-based only, U = unresolved. Multiple K codes indicate combined sources, not a ranking of theoretical maturity.

Table 26: Layout generation and spatial control in graphic design.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **LayoutTransformer** (Gupta et al., 2021)<br>G01; AI; R1, P0<br>F:S; K:E | Element boxes and labels. Layout distribution and completion. | Autoregressive likelihood fitting. |
| **LayoutDM** (Inoue et al., 2023)<br>G02; AI; R2, P2<br>F:S; K:R | Elements with operational pairwise constraints. Size / position relation satisfaction. | Denoising; relation energies guide sampling. |
| **LayoutFormer++** (Jiang et al., 2023)<br>G03; AI; R2, P2<br>F:S; K:R | Serialized element and relation constraints. Constraint satisfaction. | Likelihood fitting; restricted decoding. |
| **LayoutNUWA** (Tang et al., 2024b)<br>G04; AI; R1, P1<br>F:S; K:A | Layout attributes in editable code. Alignment and overlap, evaluated. | Reference-code instruction tuning. |
| **PosterLLaVa** (Yang et al., 2024b)<br>G05; PP; R1, P1<br>F:B; K:A+R | Poster elements as JSON boxes. Saliency-aware placement, evaluated. | Visual instruction tuning to reference layouts. |
| **LaySPA** (Li et al., 2026b)<br>G06; AI; R2, P3<br>F:B; K:A+R | Explicit element environment and relations. Spacing, alignment and text-underlay pairing. | GRPO updates the LLM layout generator. |
| **LIDA** (Dibia, 2023)<br>G07; AI; R2, P2<br>F:B; K:A+R | Data goals and executable chart code. Executable and relevant visualizations. | LLM inference, execution and selection. |

Table 27: Intermediate specifications for visualization, diagrams, and interface evaluation.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **ChartGPT** (Tian et al., 2025)<br>G08; HG; R2, P2<br>F:B; K:R | Data transformations and chart specifications. Valid data fields and chart encodings. | Specification SFT; inference validity filtering. |
| **Data Formulator 2** (Wang et al., 2025)<br>G09; HG; R2, P2<br>F:D; K:R | Data transformations and encoding specifications. Iterative analytical task completion. | User and LLM revise data / code; no new weights. |
| **AutomaTikZ** (Belouadi et al., 2024)<br>G10; AI; R2, P2<br>F:S; K:R | Executable TikZ programs. Compilable scientific vector graphics. | Code SFT; compiler-guided retry. |
| **DiagramAgent** (Wei et al., 2025)<br>G11; AI; R2, P2<br>F:S; K:A+R | Diagram plans and executable diagram code. Completeness and compilability. | Code-agent SFT; inference checking. |

*Continued on the next page.*

Table 27: Intermediate specifications for visualization, diagrams, and interface evaluation (continued).

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **AutoFigure** (Zhu et al., 2026)<br>G12; AI; R2, P2<br>F:B; K:A+R | Entity-relation blueprint and rendered diagram. Scientific fidelity and diagram clarity. | Inference critique and refinement. |
| **UIClip** (Wu et al., 2024a)<br>G13; HG; R0, P3<br>F:D; K:P+A+J | Screenshots and text assessments. Contrast, repetition, alignment, proximity. | CLIP evaluator learns perturbations / preferences. |

Table 28: Interface generation and professional design evaluation.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **UICoder** (Wu et al., 2024b)<br>G14; AI; R2, P3<br>F:S; K:R | SwiftUI code and rendered interfaces. Compilation and prompt relevance. | Filtered SFT and DPO update code generator. |
| **IDEA-Bench** (Liang et al., 2025)<br>E01; AI; Context<br>F:B; K:R+J | Images and task-specific settings. Professional task completion and quality. | Benchmark; no generator update. Not in coordinate map. |
| **Q-Judger** (Li et al., 2026a)<br>E02; PP; R0, P3<br>F:B; K:R+J | Images and explicit rubric text. 56 professional image-evaluation facets. | Supervised Q-Judger evaluator training. |
| **ArtiMuse** (Cao et al., 2026)<br>E03; AI; R0, P3<br>F:D; K:J | Images with aesthetic analyses. Eight aesthetic attributes and holistic judgment. | Expert-supervised aesthetic evaluator. |
| **GDB-Reward** (Deganutti et al., 2026)<br>E04; PP; R1, P3<br>F:B; K:A+R+J | Layout metadata, prompts and images. Text / color / layout reference fidelity. | GRPO updates prompt policy; FLUX remains frozen. |
| **PosterReward** (Lai et al., 2026)<br>E05; AI; R0, P3<br>F:B; K:A+J | Poster images and learned textual judgments. Typography, composition and preferences. | Reward learning; downstream generator RL adaptation. |

Table 29: CAD, engineering performance, and garment assembly.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **DeepCAD** (Wu et al., 2021)<br>P01; AI; R2, P1<br>F:S; K:R | Sketch / extrusion command sequences. CAD execution validity. | Command and parameter fitting. |
| **BrepGen** (Xu et al., 2024)<br>P02; HG; R2, P2<br>F:S; K:R | Face-edge-vertex B-rep structure. Geometrically valid assembly. | Diffusion fitting; geometry repaired at inference. |
| **Text2CAD** (Khan et al., 2024)<br>P03; AI; R2, P1<br>F:B; K:R+J | Text-conditioned CAD programs. Valid execution and prompt correspondence. | Conditional CAD sequence fitting. |
| **TopoDiff** (Mazé & Ahmed, 2023)<br>P04; AI; R1, P2<br>F:B; K:R+M | Material occupancy and load fields. Compliance and absence of floating material. | Surrogate gradients guide diffusion sampling. |

*Continued on the next page.*

Table 29: CAD, engineering performance, and garment assembly (continued).

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **MO-PaDGAN** (Chen & Ahmed, 2021)<br>P05; IX; R1, P3<br>F:B; K:A+M | Parametric engineering shapes. Performance and diversity. | Additional objectives train GAN generator. |
| **DressCode** (He et al., 2024)<br>P06; HG; R2, P1<br>F:S; K:R+J | Panels and sewing connections. Draping and prompt correspondence, evaluated. | Autoregressive pattern fitting. |
| **SewingLDM** (Liu et al., 2025b)<br>P07; AI; R2, P3<br>F:B; K:R+J | Panels and stitch connections. Panel / stitch integrity. | Integrity losses shape learned representation. |

Table 30: Garments, textiles, packaging, and engineering-design metrics.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **AIpparel** (Nakayama et al., 2025)<br>P08; AI; R2, P0<br>F:S; K:E | Multimodal sewing-pattern tokens. Paired panel and stitch reconstruction. | Supervised multimodal generation / editing. |
| **TextileDiffusion** (Karagoz et al., 2023)<br>P09; DS; R0, P0<br>F:S; K:E | Textile-pattern images. Appearance and text similarity. | Image-domain diffusion adaptation. |
| **Knit-FLUX** (Liu et al., 2025c)<br>P10; IX; R0, P1<br>F:S; K:A+J | Rendered knit images. Knit-unit appearance and yarn coverage, evaluated. | Image-domain LoRA adaptation. |
| **FunctionalTextile** (Tang et al., 2026)<br>P11; DS; R2, P3<br>F:B; K:R+M | Reported yarn / interlacing fields. Reported porosity, stiffness and conductivity. | Reported functional-surrogate loss trains generator. |
| **KumisPackaging** (Zhao et al., 2026)<br>P12; IX; R1, P2<br>F:D; K:R+J | Cultural elements and shape-grammar variants. Cultural communication and packaging preference. | Human prompting and candidate selection. |
| **CraftBeerPackaging** (Chen & Jang, 2026)<br>P13; DS; Abstract only<br>F:D; K:U | Packaging images from existing tools. Visual and brand-expression assessment. | Tool comparison; abstract-level evidence. Not in coordinate map. |
| **Beyond Statistical Similarity** (Regenwetter et al., 2023)<br>P14; DS; Context<br>F:B; K:A+R+M | Cross-engineering evaluation taxonomy. Feasibility, performance, novelty and conditioning. | Comparative metric study; contextual reference. Not in coordinate map. |

Table 31: Urban, landscape, and story visualization.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **UrbanGAN** (Fedorova, 2021)<br>S01; DS; R1, P1<br>F:S; K:A+R | Urban-block rasters and parcel context. Building spacing and urban morphology, evaluated. | Conditional adversarial fitting. |
| **DRL-Urban** (Zheng et al., 2023)<br>S02; IX; R2, P3<br>F:D; K:P+R | Parcel-road-intersection graph. Service access, ecology and road efficiency. | PPO trains a planning policy from scratch. |
| **PlantoGraphy** (Huang et al., 2024)<br>S03; HG; R2, P2<br>F:B; K:R+J | Scene graph and plant-layout constraints. User-specified landscape composition. | Interactive graph planning; rendering LoRA. |
| **GenFusion-Urban** (Gao et al., 2025)<br>S04; DS; Abstract only<br>F:B; K:R+M | Reported urban layout / performance pipeline. Urban building performance. | Reported diffusion + surrogate + NSGA-II. Not in coordinate map. |
| **TaleCrafter** (Gong et al., 2023)<br>S05; HG; R1, P2<br>F:S; K:R | Character attributes, layouts and story sequence. Character consistency and story visualization. | Adapted rendering plus inference control. |

Table 32: Cinematic organization and playable levels.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **CineVerse** (Phung et al., 2026)<br>S06; AI; R2, P2<br>F:B; K:P+R | Structured cinematic shot plans. Cinematic composition and continuity. | Plan-guided inference; keyframe fitting. |
| **STAGE** (Zhang et al., 2026b)<br>S07; AI; R2, P3<br>F:B; K:A+J | Intra-shot and inter-shot frame pairs. Cinematic transition preference. | DPO updates frame-pair predictor; video model frozen. |
| **ShotPlan** (Guo et al., 2026)<br>S08; PP; R2, P1<br>F:B; K:R+J | Explicit shot / cut planning tokens. Shot controllability and consistency, evaluated. | Conditional flow matching; no separate planning loss. |
| **Cont-PCGRL** (Earle et al., 2021)<br>S09; AI; R2, P3<br>F:D; K:A+R | Grid regions, paths and edit actions. Target difficulty / path and region objectives. | From-scratch goal-conditioned PPO. |
| **DiffusionCraft** (Dai et al., 2024)<br>S10; AI; R1, P1<br>F:B; K:R | Level tiles and local spatial patterns. A-star playability, evaluated. | Diffusion fitting from one level. |

Table 33: Controllable levels and exhibition layouts.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **PCGRL-Jax** (Earle et al., 2024)<br>S11; AI; R2, P3<br>F:D; K:A+R | Grid edits, goals and fixed elements. Path / region control and generalization. | From-scratch PPO with scalable environments. |
| **IPCGRL** (Baek et al., 2025)<br>S12; AI; R2, P3<br>F:D; K:A+R | Language-derived tasks and level edits. Instruction-specified paths and regions. | Task encoding plus reward-trained PPO. |
| **Museum-CGAN** (Tang et al., 2024a)<br>S13; IX; R1, P1<br>F:B; K:J | Museum zones and rendered plans. Spatial / visual qualities, evaluated. | Adversarial and perceptual image fitting. |
| **Rule-Exhibit** (Xu et al., 2025)<br>S14; HG; R2, P3<br>F:B; K:A+R | Hall connectivity and exhibit placement. Connectivity, saliency and exhibit importance. | Design rules generate supervised training labels. |
| **Museum-LoRA-Compare** (Li et al., 2026c)<br>S15; DS; Abstract only<br>F:B; K:T+J | Generated museum-plan images. Spatial connectivity, evaluated. | LoRA / GAN comparison; abstract-level evidence. Not in coordinate map. |

Table 34: Floorplan synthesis and design-directed learning.

| Study / source | Representation and criterion | Where the criterion acts |
|---|---|---|
| **House-GAN++** (Nauata et al., 2021)<br>F01; AI; R2, P2<br>F:S; K:R | Room graph and room masks. Graph compatibility and geometric quality. | GAN training; schedule meta-optimization. |
| **HouseDiffusion** (Shabani et al., 2023)<br>F02; AI; R2, P1<br>F:S; K:R | Graph-conditioned vector polygons. Geometry and compatibility, evaluated. | Noise and binary-coordinate regression. |
| **ChatHouseDiffusion** (Qin et al., 2024)<br>F03; PP; R2, P2<br>F:S; K:R | Language-derived relations and raster plan. User conditions and controlled editing. | Conditional denoising and inference editing. |
| **GSDiff** (Hu et al., 2025a)<br>F04; AI; R2, P3<br>F:S; K:R | Boundary graph and vector geometry. Alignment and edge validity. | Native geometric losses train generator. |
| **FloorPlan-LLaMa** (Yin et al., 2025)<br>F05; AI; R0, P3<br>F:B; K:J | Image tokens; image-based preference evaluator. Expert functionality, flow and overall judgment. | PPO updates autoregressive image-token generator. |
| **Floorplan RLVR** (Lara et al., 2026)<br>F06; AI; R2, P3<br>F:S; K:R | JSON polygons and reconstructed relations. Connectivity-graph and area requirements. | GRPO updates autoregressive generator. |
| **GreenPlanner** (Zeng et al., 2026)<br>F07; AI; R2, P3<br>F:B; K:P+R+M | Conditioned vector / topology representation. Learned energy and feasibility criteria. | Gradient feedback adapts normalizing-flow generator. |

# G Design Theory Translation: Principles for Model Improvement

Building on the floorplan results and cross-domain review, we propose a workflow for translating design theory into model improvement in other tasks. The review identifies three connections: design progress needs purpose-specific evidence (RQ1, Section F.2); a computable criterion needs a justified domain meaning (RQ2, Section F.3); and feedback needs a route into the generator's own decisions (RQ3, Section F.4). Their common prerequisite is an operational account of how design organization serves its purpose. We therefore begin with *GenAI-native theory construction*, then instantiate its representation, empirical reference, and learning interface, derive design principles from the floorplan evidence, and examine how agent-assisted implementation can extend the procedure across tasks (Figure 21).

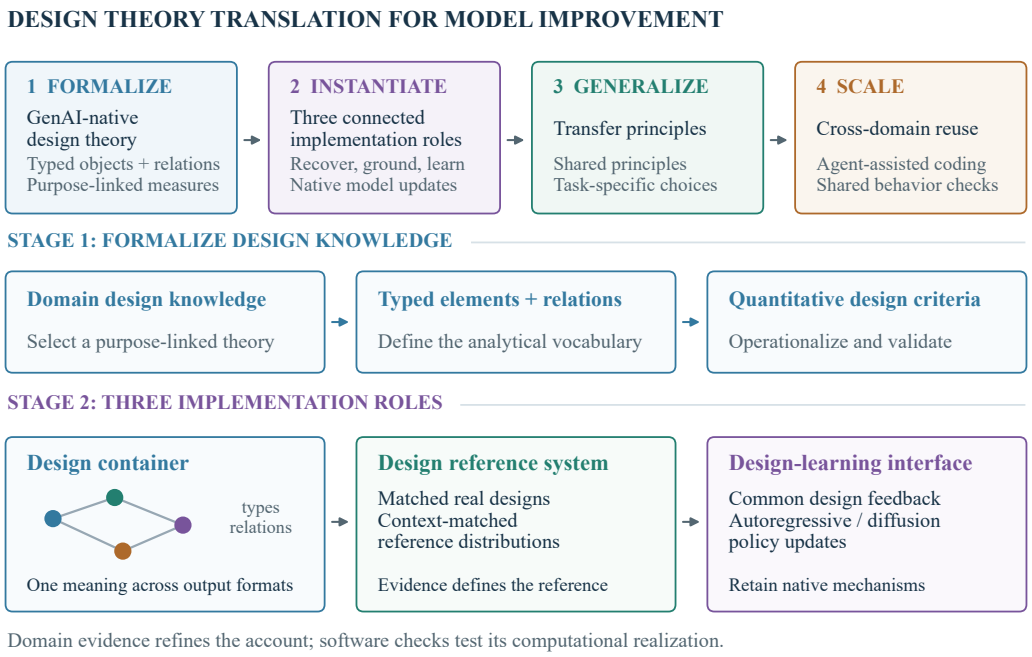



Figure 21: **A proposed workflow for design theory translation.** Informed by the floorplan study, the overview connects formalization, implementation, and cross-domain extension. Stage 1 defines typed analytical objects and quantitative criteria grounded in domain meaning. Stage 2 recovers the design container from artifacts, grounds measurements in empirical references, and connects shared feedback to native model updates.

## G.1 Generative AI-Native Design Theory Construction

**The translation challenge.** Design disciplines offer extensive theories and principles of how artifacts serve their purposes. The GenAI-era challenge is to translate this knowledge into model-readable representations and quantitative measurements that guide generator learning. Where analytical objects or measures remain unspecified, this requires constructing a design container and its evaluator. This motivates *Generative AI-native design theories*: operational accounts developed or reformulated around the design elements, typed relations, contextual purposes, and quantitative properties that a generative learning process can use. Here, "native" describes compatibility with that process, rather than the historical origin of the theory.

Space syntax supplies a particularly useful precedent. Long before contemporary GenAI, it connected a theory of spatial organization to graph representations and quantitative measures of accessibility (Hillier & Hanson, 1984; Hillier, 1996). Its analytical objects and measures can therefore support a design container without first inventing a new architectural criterion. SSIO supplies the recovery operator; residential corpora supply empirical references; SSPT supplies the learning connection. The

reusable lesson is this development pattern—*design purpose, relational representation, quantitative analysis, and evidence of design consequences*—rather than applying spatial integration to unrelated tasks.

**Constructing a relational account.** For a domain without a ready operational account, begin with the existing theory that most directly explains the intended design purpose and admits analysis of relations. Its concepts determine the container's vocabulary: what counts as an element, which functional types matter, and which directed, spatial, temporal, or functional relations distinguish successful organization. A typed, attributed graph provides a common form; geometry, material, time, and user context remain attached where the theory requires them. These types distinguish, for example, a label from a depicted process, a control from a navigation destination, or a service space from a public space. Merely connecting visually adjacent elements would miss these meanings.

Next, translate the theory's claims into measurable properties of this container. Retain established formulas where available; otherwise state the proposed operational definition and the controlled design changes it should distinguish. A grouping principle might motivate a discrepancy between intended groups and represented regions; a task-performance model might predict the time of an executable interaction sequence. These are different kinds of evidence: the first introduces a theory-inspired metric, whereas the second implements an existing quantitative model. Domain tasks, professional judgments, or observed use establish whether the property captures the intended benefit. Real-design distributions then determine contextual reference levels.

**Designing for the generative process.** The operational account should remain stable across the formats in which a generator expresses the same design. It must specify how complete outputs expose the required elements and relations, how unsupported outputs are recognized, and how alternative organizations are compared under a common brief. These requirements support both autoregressive and diffusion generators: a terminal design property can guide their different native updates without requiring a differentiable graph-recovery procedure. Generative experiments can also expose missing distinctions in the account, motivating revised types, relations, or measurements supported by domain evidence.

The fourteen-domain map in Section G.7 distinguishes reusable quantitative foundations from theories requiring further operationalization and from complementary standards or physical models. A domain may contain all three for different purposes. Where a ready-made theory is unavailable, formalize the relevant traditional knowledge, combine compatible accounts where necessary, and test how the proposed organization relates to design outcomes. The following three roles connect this analytical foundation to model improvement.

### G.2 The Design Container: A Theory-Defined Intermediate Representation

The preceding theory construction specifies *what the design container means*; implementation specifies *how it is recovered from actual artifacts*. A useful common form is a typed, attributed graph $z_d = (V_d, E_d, \tau_d, a_d, c)$: $V_d$ contains design elements, $E_d$ their relations, $\tau_d$ assigns element and relation types, $a_d$ stores attributes, and $c$ records the design context. Higher-order relations can be represented by relation nodes. Spaces and access, information units and reading order, garment panels and seams, or components and functional interfaces instantiate different semantics within this form.

We call this analytical interface a *design container*: a theory-defined intermediate representation of the objects, relations, attributes, and context needed to compute design properties. Recovery operators map native outputs into this representation without requiring generators to share an output format or synthesis mechanism.

Compiler intermediate representations provide a useful analogy: LLVM and MLIR support shared analyses across different source languages through explicit semantics (Lattner & Adve, 2004; Lattner et al., 2021; LLVM Project, 2026). Here, $\mathcal{A}_{d,m}$ recovers a domain representation $z = \mathcal{A}_{d,m}(x, c)$ from artifact $x$ and context $c$, and theory-defined measure $\mathcal{M}_d$ computes $s = \mathcal{M}_d(z, c)$. Likewise, cross-model design comparison requires consistent meanings for entities and relations, specified recovery rules for each format, and common measurement and aggregation rules.

SSIO instantiates this container with convex-space access graphs and integration. The GSDiff case shows why construction must distinguish recoverable information from missing relations (Section 5.5; Appendix E.6): a room-boundary graph does not determine access. When completion supplies those relations, its recovery fidelity and the generator's subsequent improvement require separate evaluation. The container thus specifies both a shared analytical meaning and the information needed to support it.

### G.3 The Design Reference System: Theory, Evidence, and Evaluation

A *design reference system* combines theory-defined properties with contextual evidence from real designs. Intended use defines the design objective; theory makes it measurable; real designs establish context-matched references. Recurrence identifies organizational patterns whose desirability is interpreted through that purpose. Comparisons preserve relevant requirements, scale, functions, users, and environment, recovering the same design container from references and model outputs.

Multiple corpora separate robust regularities from collection-specific conventions. Theory, task performance, simulation, or professional assessment can corroborate their interpretation. The result is an empirically grounded reference without imposing a single numerical template on multiple valid solutions.

The floorplan findings illustrate this division. Public-first organization motivates the functional property, while corpus-dependent distributions require matched numerical references. RPLAN-SSI evaluates population-level profile alignment and within-plan Public anchoring; the matched generator–corpus analysis separately diagnoses the weaker functional hierarchy produced by a generator.

The same construction can connect three levels elsewhere: verify objects and relations, evaluate an individual artifact against its purpose, and compare population distributions, coverage, and variability with matched real designs. A benchmark should publish recovery rules, reference splits, contextual matching, metrics, coverage, and uncertainty. Recovery failures remain visible so score changes cannot be detached from measurement coverage.

### G.4 The Shared Design-Learning Interface: Feedback and Native Updates

A *shared design-learning interface* uses common measurements of completed designs to guide native policy updates across generators. Reference distributions calibrate and evaluate the objective; artifact-level feedback directs model updates. Shared design meaning connects the different model action spaces.

For an *autoregressive generator*, actions are sequence decisions. A complete artifact can be decoded into its design container and measured to supply rewards for policy-gradient updates or rankings for preference optimization (Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). Tokenization, likelihoods, and action masks remain native to the sequence model.

For a *diffusion generator*, actions are stochastic reverse transitions. DDPO supplies this policy-learning view (Black et al., 2024); each backbone's reverse variables, transition probabilities, conditioning, masks, and variances determine its policy update. SSPT demonstrates these bindings across five floorplan backbones, and its HouseDiffusion support ablation shows that update placement matters under a fixed transition budget.

The division is precise: share design semantics and measurement, but bind learning to the native generative policy. SSPT provides the diffusion instantiation; an autoregressive implementation would use the same feedback meaning through sequence-specific updates.

### G.5 Retaining Required Quality During Adaptation

First, specify how much output quality may deteriorate and which validity requirements must hold. Use development experiments to tune update strength and control reward shortcuts. On validation conditions, discard checkpoints that violate these requirements, then select the remaining checkpoint with the best design performance. Finally, evaluate the selected checkpoint on independent test conditions, reporting output quality and design performance separately.

HouseDiffusion illustrates this trade-off. Retrospective calibration from frozen Base evaluations establishes a native-quality reference range (Appendix C.3). The selected model remains within this range while IPD decreases by 27.0% and median PIA increases by 41.9%; FID and Compatibility increase relative to matched Base. Area control reduces the native-quality cost of configurational feedback. Sensitivity analysis shows how changing the quality requirement changes checkpoint choice (Appendix E.4.6).

Other domains require their own gate—rendering validity, geometric consistency, program execution, or simulation feasibility. Limited adaptation, native-objective regularization, and reference-policy control offer candidate preservation mechanisms; independent checkpoint tests measure their effectiveness.

### G.6 From Floorplan Evidence to a Reusable Research Procedure

The proposed construction procedure begins before metric selection: identify the design purpose, then construct a GenAI-native analytical account when one is not already available. Recover its design container, ground its measurements in matched empirical references, and connect feedback to native model updates. Independent design outcomes and retained output quality evaluate whether the resulting change constitutes improvement. These steps distinguish theoretical meaning, computational realization, and empirical success.

The floorplan evidence connects each step to a test: cross-corpus comparisons distinguish regularities from context-dependent references; matched cases separate organization from link count; ablations isolate update placement and shortcut control; and GSDiff separates relation recovery from learning through the recovered structure. The domain sketches below translate these tests into construction questions for other tasks.

### G.7 A Fourteen-Domain Map of Construction Routes

Readiness belongs to a specified design purpose, rather than an entire discipline. The fourteen sketches identify the available analytical foundation and the construction still required. Quantitative theories can supply a container and measures directly; explanatory theories and principles require operationalization; standards and physical models can supply complementary formal objects or tests. Keeping these sources distinct preserves the review's provenance analysis (Appendix F.3). References and generated outputs then use common recovery rules, with controlled changes and domain outcomes testing the meaning of the resulting scores.

Table 35 summarizes these entry points. Existing quantitative accounts are especially useful for spatial access, expert task execution, requirement-defined product design, and specified graphical encodings. For communication, continuity, and apparel needs, the selected accounts require a further bridge from explanation to a task-specific measure. Formal notations, symmetry, folding mechanics, and graph grammars provide additional construction resources whose contribution should be stated in their own terms. The detailed sketches identify both inherited analytical content and the proposed extension.

#### G.7.1 Communication and Interaction

Here, organization primarily serves interpretation or task completion. Typography, data mappings, temporal order, and executable transitions belong in the container whenever they carry design meaning; references must be matched by content, audience, and task.

**Poster and graphic layout.** Common-region grouping is an established perceptual principle explaining association (Palmer, 1992), while learned layout analyses already quantify importance, alignment, and grouping (O'Donovan et al., 2014). Text, images, visible regions, typography, and intended priority form a candidate container. Priority inversions and disagreement between intended groups and region membership are proposed task-specific measures, rather than formulas inherited from the perceptual principle. Brief-matched posters and message-identification or recall tasks can ground and validate them.

Table 35: **Construction entry points across fourteen design categories.** Groups describe the selected knowledge sources and purposes; the paragraphs below provide their evidence and operator proposals.

| Design category | Available foundation | Construction focus |
|---|---|---|
| *Existing quantitative or formal analytical accounts* | | |
| Floorplans / interiors | Space syntax | Configuration recovery and grounding |
| Urban / landscape | Space syntax; cooling models | Purpose-specific network or climate inputs |
| UI / interaction | Keystroke-Level Model | Executable methods and calibrated timings |
| Products / CAD | Axiomatic Design | Functional dependencies and success evidence |
| Visualization / maps | Expressiveness; projection theory | Task-matched encodings and distortion |
| Exhibition / wayfinding | Access and cross-visibility analysis | Visitor goals and interpretive outcomes |
| *Explanatory theories and principles requiring operationalization* | | |
| Posters / graphics | Common-region grouping | Intended versus perceived grouping |
| Editorial | Discourse relations; spatial contiguity | Content-to-layout correspondence |
| Cinema / storyboards | Attentional continuity theory | Cut-specific continuity measurements |
| Garments | Functional–expressive–aesthetic needs | Purpose-specific fit and movement criteria |
| *Complementary formal methods, standards, and physical models* | | |
| Scientific diagrams | SBGN typed notation | Semantic conformance and comprehension |
| Textiles / patterns | Symmetry; textile simulation | Repeat structure or physical performance |
| Packaging | Fold kinematics; construction/test standards | Assembly, use, and protection objectives |
| Game levels | Mission–space grammars; MDA | Progression constraints and player outcomes |

**Editorial design.** Rhetorical Structure Theory supplies a theory of discourse functions (Mann & Thompson, 1987), and spatial contiguity relates text–figure placement to comprehension (Mayer, 1999). Document hierarchy, reading order, references, page geometry, and language direction provide the analytical objects. Reading-order violations, broken references, and text–figure separation can be evaluated against complete publications with equivalent content, then checked through retrieval time and comprehension.

**Scientific and mechanism diagrams.** Congruence and apprehension connect graphical organization to conceptual understanding (Tversky et al., 2002); Systems Biology Graphical Notation (SBGN) supplies typed semantic and layout rules for biological mechanisms (Rougny et al., 2019). Entities, states, processes, compartments, and directed interactions form a domain container. Conformance checks inherit the standard; a weighted violation score is an additional operational choice. Expert mechanism specifications and process-tracing tasks separately evaluate scientific correctness and comprehension; other sciences require their own ontologies.

**Data visualization and maps.** Graphical expressiveness formalizes faithful encoding, while effectiveness concerns task-dependent perceptual readability (Mackinlay, 1986); projection theory quantifies geographic distortion (Snyder, 1987). Typed fields, marks, scales, geographic features, units, and projection parameters expose these properties. Encoding checks and projection equations provide existing analytical foundations. A combined perceptual score requires its own task assumptions and evidence, rather than treating all effectiveness rankings as established quantitative laws.

**UI and interaction design.** The Keystroke-Level Model (KLM) predicts time for a specified expert, routine, error-free method (Card et al., 1980). Ordered operations, devices, mental preparation, and

response delays therefore supply a directly quantitative container. Screens, controls, and executable transitions add task structure: reachability is a separate property, and error rates require observed interaction. Method recovery and calibrated timings connect the model to generated interfaces; broader usability and accessibility need corresponding users and outcome evidence.

**Cinematic composition and storyboards.** The Attentional Theory of Cinematic Continuity explains how attention and perceptual expectations support continuity across cuts (Smith, 2012). Shots, events, characters, trajectories, gaze targets, and cut types expose that organization. Action-phase mismatch and target displacement are candidate operationalizations, not ready-made scores supplied by the theory. Their interpretation requires cut-specific continuity goals, timed sequences, and gaze or comprehension evidence, including cases where discontinuity is intentional.

**Exhibition and wayfinding design.** Exhibition studies distinguish accessibility from directed cross-visibility (Peponis et al., 2004), while interpretive guidance links information to visible site features (National Park Service, 2019). Separate movement and visibility graphs can represent exhibits, paths, signs, destinations, occlusion, and thematic links. Route success, engagement, and interpretive understanding provide stronger grounding than visibility alone.

### G.7.2 Spatial and Physical Design

Physical relations must be interpreted under use conditions: loads for parts, bodies and materials for garments, climate for landscapes, and movement rules for playable levels. Reference artifacts provide organizational context; simulation, testing, or observed use supplies performance evidence.

**Floorplans and interiors.** Space syntax formalizes access organization through spaces, connections, and integration (Hillier & Hanson, 1984; Hillier, 1996). SSIO already recovers this structure and relates functional integration to residential reference distributions. An interior extension could add furniture, activity locations, and mobility envelopes, grounding activity reachability and route length in furnished plans and observed use.

**Urban and landscape design.** Space syntax relates street-network configuration to pedestrian movement and land-use attraction (Hillier et al., 1993). Street segments, crossings, destinations, transit stops, and pedestrian evidence would support accessibility and service coverage. A separate performance-modeling branch could use InVEST to estimate urban heat mitigation from land cover, shade, albedo, and evapotranspiration under reference climate conditions (Bosch et al., 2021); the planning brief determines which branch defines improvement.

**Industrial products and mechanical CAD.** Product architecture maps functions to components and interfaces (Ulrich, 1995). Axiomatic Design adds a quantitative foundation: functional requirements and design parameters form a dependency matrix, while independence and information axioms analyze their coupling and probability of satisfying requirements (Suh, 1996). A container can preserve these typed dependencies alongside parts, joints, materials, loads, and tolerances. Recovering functional assignments and estimating requirement-success probabilities are the adaptation tasks; mechanical simulation supplies purpose-specific performance evidence (Autodesk, n.d.).

**Garments.** The functional, expressive, and aesthetic framework organizes apparel needs but does not itself provide quantitative fit or comfort formulas (Lamb & Kallal, 1992). Panels, seams, bodies, materials, and activity provide a candidate container; GarmentCodeData supplies pattern representations and draping checks (Korosteleva et al., 2024). Pattern feasibility is already computable. Connecting ease, fit, movement, or expression to the framework requires matched bodies, fabrics, activities, and wearer evidence.

**Textiles and surface patterns.** TexGen constructs yarn geometry for mechanical and permeability simulations (Brown, 2022), whereas plane-group symmetry formalizes motif repetition (Meier, 1984). These purposes require distinct containers: yarn paths, interlacing, contacts, material properties, and loads for physical performance; motifs, repeat boundaries, and symmetry for surface organization. Manufactured swatches or repeat-ready designs provide the corresponding references.

**Packaging.** Fold kinematics offers a narrow quantitative foundation: faces and hinge relations can determine whether a specified package admits the required folding motion (Balkcom et al., 2009). FEFCO supplies a construction vocabulary and ISTA supplies distribution-performance tests (FEFCO, 2022; International Safe Transit Association, n.d.); these are complementary standards. Panels, folds, tabs, closures, inserts, and product geometry expose assembly and retention. Protection and opening objectives additionally require material, product, shipment, and use evidence.

**Game levels.** Mission–space graph grammars provide a formal route from progression structure to playable space (Dormans, 2010). Regions, hazards, resources, gates, prerequisites, and state-dependent transitions form a container for progression checks. MDA relates mechanics and dynamics to player experience (Hunicke et al., 2004), while expressive-range analysis characterizes generator distributions (Smith & Whitehead, 2010). Reachability and prerequisite consistency are directly testable; challenge and experiential improvement require a declared player model and play traces.

### G.8 Agent-Assisted Translation and Worked Examples

Three examples connect design goals to computable feedback: task completion and predicted effort in UI using KLM (Card et al., 1980); typed process roles in scientific diagrams using SBGN (Rougny et al., 2019); and intended poster grouping using an authored metric motivated by common-region perception (Palmer, 1992). Figure 22 shows their analytical objects and controlled changes. The examples show how recovered design structure supports task-specific evaluation and feedback-guided editing, without generator training or human-performance evaluation.

**UI: completing the task before reducing effort.** Three authored tasks—saving a document, formatting a selection, and exporting a PDF—each have a menu method, a shortcut, and an incomplete alternative. An executable state model verifies the requested result independently of the KLM time calculation. For saving, supplied operation times predict 4.4 seconds for the menu method and 1.8 for the shortcut; merely opening File takes 3.1 seconds but leaves the document unsaved. Across nine methods, six complete their tasks and three incomplete alternatives are excluded from time-based improvement. The reference is a complete method for the same task, not a pooled distribution of unrelated interfaces.

**Scientific diagrams: recovering the meaning of a connection.** Three official libSBGN example maps—glycolysis, MAPK, and insulin-like growth-factor signaling—supply 19 compatible irreversible process neighborhoods. Removing an entire consumption or production role is detected in all 38 paired changes; all 38 wrong-direction or wrong-type variants are rejected. However, 16 single-arc deletions leave another arc fulfilling the same role and do not change the score. Thus, role completeness differs from preserving every reaction participant: the latter requires a richer, task-specific account of the process, not simply more connectivity.

**Feedback-guided poster repair.** Seven authored event briefs use six content elements in two or three panels: three development briefs and four evaluation briefs, with each brief's variants kept together. We compare no editing, geometry-only editing, and grouping-feedback editing. Both editors evaluate 22 candidates per round for four rounds (88 evaluations per editor and brief), using the same operations: no change, all pair swaps, and individual grid snaps. Geometry minimizes grid displacement and overlap; grouping feedback first minimizes pairwise disagreement between intended groups and displayed regions, then geometry. Equal scores retain the current design.

Table 36: **Poster repair under a shared editing budget.** Means over four authored evaluation briefs. Named-panel accuracy is evaluated separately from the grouping objective.

| Selection criterion | Group disagreement ↓ | Grid error ↓ | Named panel ↑ |
|---|---|---|---|
| No editing | 0.350 | 0.463 | 58.3% |
| Geometry only | 0.350 | 0.093 | 58.3% |
| Grouping feedback | 0.083 | 0.278 | 79.2% |

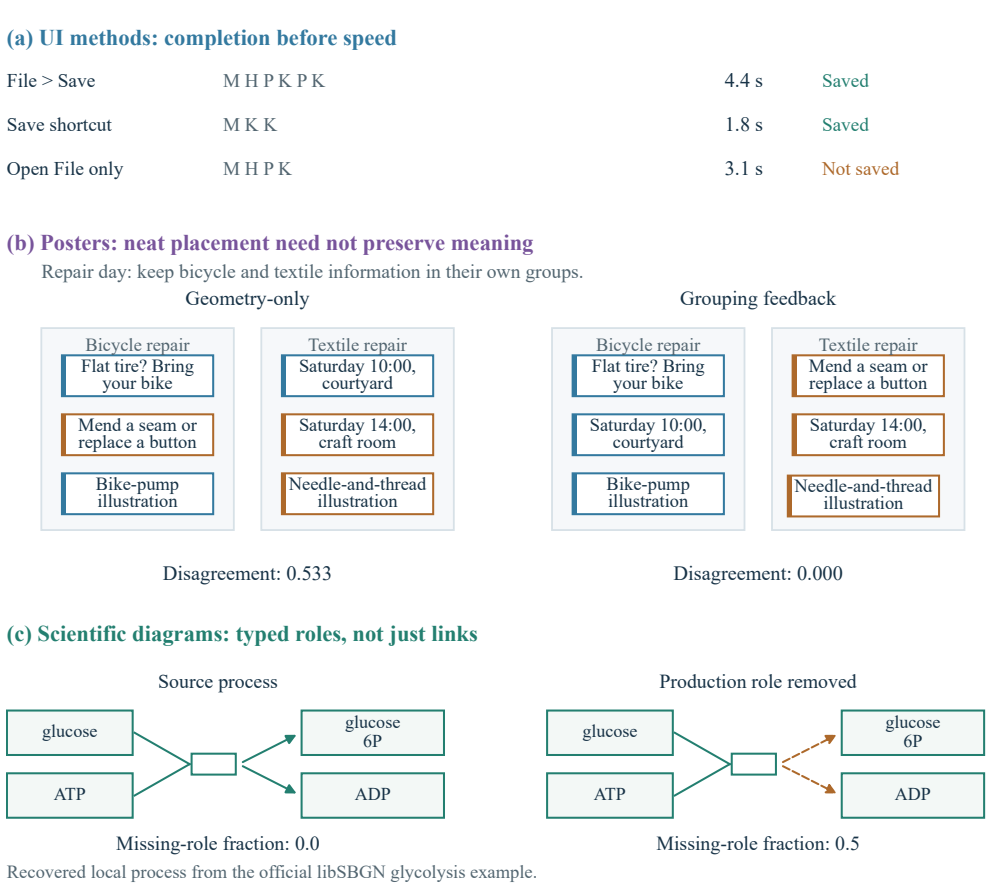



Figure 22: **Worked theory-translation examples.** (a) Task completion qualifies KLM time comparisons; M/H/P/K denote mental preparation, homing, pointing, and keystrokes. (b) Two editors start from the same development poster. Colors identify intended groups; image contents appear as labels in the recovered container. (c) A process recovered from the libSBGN glycolysis example; dashed arcs mark deleted production relations.

Grouping feedback reduces mean disagreement from 0.350 to 0.083, whereas geometry-only editing improves alignment without repairing groups (Table 36). Both preserve content and avoid overlap or containment errors; grouping repair leaves more alignment error under the shared budget. The library poster reaches zero disagreement with only one-third of its elements in their intended named panels, distinguishing grouping from semantic assignment. The clinic's groups of four and two cannot fit two three-slot panels: exhaustive enumeration confirms that disagreement of $1/3$ is unavoidable without changing the template. Effective feedback therefore needs panel semantics, and the edit space must permit the intended organization.

**A reusable implementation workflow.** Researchers specify purposes, objects, relations, and measurement sources; guided agents implement design containers, task-matched references, and feedback for controlled editing. Across examples, equivalent formats and renamed elements preserve recovered organization and scores; poster grouping is also invariant to whole-layout translation and scaling. Checks cover finite scores, context matching, and reference/test separation. Under `framework/`, instructions are in `skills/design-theory-translation`; code, sources, and results in `extensions/design_theory_translation/worked_validation`.

The cases identify three requirements for extension: completion before efficiency, semantically faithful relations, and expressive edit spaces. Trained-generator studies can test resulting design gains and reductions in expert effort.